UNIVERSITÀ DEGLI STUDI DI GENOVA

PHD SCHOOL IN EXPERIMENTAL MEDICINE

CURRICULUM BIOCHEMISTRY

XXXII CYCLE

AA 2019/2020

Study and characterization of Glycosyltransferases from Paramecium bursaria Chlorella virus – 1

CANDIDATE: Dr. Maria Elena Laugieri TUTOR: Prof. Michela Tonetti

BIO /10

Index

Figures index

Tables index

Abbreviations

#

5mC	5-methylcytosine
6mA	N6-methyladenine

A

Ara	Arabinose
Asn	Asparagine
Asp	Aspatic acid
ATCV-1	Acanthocystis turfacea chlorella virus 1

C

CDG	Congenital Disorders of Glycosylation
CDNB	1-chloro-2,4-dinitrobenzene
CHS	chitin synthase
COPI	Coat protein complex I
CPS	capsular polysaccarydes
CV	column volumes
Cys	Cysteine

D

DNA	Deoxyribonucleic acid
Dol-P	Dolichol phosphate
Dol-P-P-GlcNAc	Dolichol pyrophosphate N-acetylglucosamine
D-Rha	D-rhamnose

E

EDTA	Ethylenediaminetetraacetic acid
ER	Endoplasmic Reticuulum
ERAD	ER-associated protein degradation
EXT	Exostosin

F

FPLC	Fast protein liquid chromatography
Fuc	Fucose

G

GDP	Guanidine diphosphate
GFAT	glutamine-fructose-6P aminotransferase
Glc	Glucose
GlcNAc	N-acetyl glucosammine
GlcNAcTs	N-acetylglucosaminyltransferases
Gln	Glutamine
Glu	Glutamic acid
Gly	Glicine
GMD	GDP-D-mannose 4,6-dehydratase
GMER	GDP-4-keto-6-deoxy-mannose epimerase/reductase
GPI	glycosylphosphatidylinositol
GSH	Gluthatione
GST	Gluthatione S- transferase
GT	glycosyltransferase
GT-A	glycosyltransferase A fold
GT-B	glycosyltransferase B fold
GT-NC	glycosyltransferase not classified

H

HAS	hyaluronan synthase
His	Histidine
HPLC	High Perfprmance Liquid Chromatography

I

IPTG	Isopropyl β-d-1-thiogalactopyranoside
ITC	some Isothermal Titration Calorimetry

K

Kd	dissociation constant
kDa	kilo dalton

L

LB	Luria Bertani
Leu	Leucine
LLO	lipid linked oligosaccharide
L-Rha	L- rhamnose

M

Man	Mannose
MW	Molecular weight

N

NAD+	Nicotinamide adenine dinucleotide
NADP+	Nicotinamide adenine dinucleotide phosphate
NADPH	Nicotinamide adenine dinucleotide phosphate reduced form
NCLDVs	Nucleo cytoplasmic large DNA viruses
NMR	Nuclear Magnetic Ressonance

O

ORF	Open Reading Frame
OST	oligosaccharyltransferase
OTase	oligosaccharyltransferase

P

PBCV -1	Paramecium bursaria Chlorella virus – 1
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PCT	pre crystallization test
PglK	ATP-dependent flippase
Phe	Pheninalanine
PI	Post Infection

R

RNA	Ribonucleic acid

S

SDS-PAGE	sodium dodecyl sulfate–polyacrylamide gel electrophoresis
Ser	Serine
SFT1	Protein transport protein
SN2	substitution nucleophilic reaction 2
Sni	substitution nucleophilic reaction 2
STD	Saturation Transfer Difference

T

Thr	Threonine
TRIS	tris(hydroxymethyl)aminomethane

U

UDP	Uridine Diphosphate
UDP-D-GlcNAc	UDP-D-N-acetylglucosamine
UGD	UDP-D-glucose dehydratase
UGDH	UDP-D-glucose dehydrogenase
UGGT1	UDP-Glc:glycoprotein glucosyltransferase
Und-PP	Undecaprenyl-pyrophosphate

V

Val	Valine

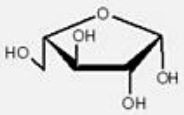
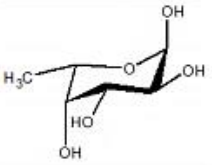
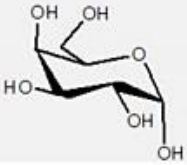
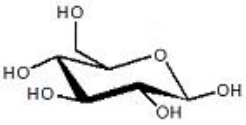
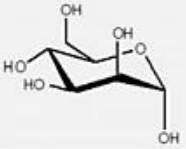
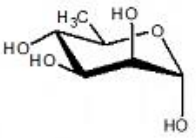
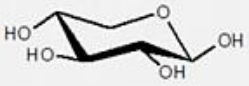
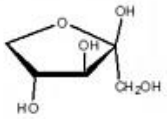
W

WT	Wild Type

X

Xyl	xylose

Monosaccharides symbols

Araf	β-L-arabinofuranose		
Fuc	α-L-fucose		
Gal	α-D-galactose		
Glc	β-D-glucose		
Man	α-D-mannose		
Rha	α-D-rhamnose		
Xyl	β-D-xylose		
Xulf	β-D-xylulose		

Amino acids

Amino acid	3-letter abbreviation
Alanine	Ala
Arginine	Arg
Asparagine	Asn
Aspartic acid	Asp
Cysteine	Cys
Glutamic acid	Glu
Glutamine	Gln
Glycine	Gly
Histidine	His
Isoleucine	Ile
Leucine	Leu
Lysine	Lys
Methionine	Met
Phenylalanine	Phe
Proline	Pro
Serine	Ser
Threonine	Thr
Tryptophan	Trp
Tyrosine	Tyr
Valine	Val

Abstract

Giant Viruses are a class of uncommon cellular parasites discovered about 30 years ago[1]. They are defined as Nucleo Cytoplasmic Large DNA viruses (NCLDVs) according to the notable viral particle dimensions (about 400nm) and the genome complexity. In addition, NCLDVs possess genes with "cell-like properties", that allow the virus to be, at least in part, independent from the host molecular mechanisms[1]. One of the most important pathways that is almost totally encoded by NCLDVs is the glycosylation. Generally, viruses use the ER/Golgi compartments of the host to glycosylate their own proteins. For NCLDVs, an almost complete system to elongate, modify and synthetize the glycoforms is set up in the viral factories, which are defined structures in the host cytoplasm.

The topic of this work was the study and the characterization of two of the six putative glycosyltransferases (GTs) from *Paramecium bursaria Chlorella virus- 1* (PBCV-1): A064R and A075L[2]. PBCV-1 possesses in fact an highly glycosylated capsid that displays uncommon glycoforms only shared by chloroviruses[3]. The identification of the glycoform structure suggest that they are probably synthetized by the virus and not by the host. These findings represented the starting point to analyse PBCV-1 genome looking for genes encoding those enzymes. In the present work, A064R is characterized by enzymatic analysis, demonstrating that it is a multidomain enzyme, with two rhamnosyltransferase activities and a methyltransferase one. A075L is also demonstrated to be a GT, by enzymatic analysis and ITC experiments. Experiments aimed also to identify the 3D structure of the protein, and to confirm its interaction with the substrate, the UDP-xylose. The solving of the 3D structure and the enzymatic characterisation are currently underway.

A064R and A075L enzymes display interesting catalytic properties that could be explored for biotechnological applications. In fact, the study of the enzymes that process glycans is a recent topic explored for the production of compounds largely used as bioactive molecules [4]. Identification of novel GTs will provide new tools that can expand the biological biodiversity of glycans as bioactive natural products, which is well known to participate in the molecules drug efficiency in terms of

pharmacokinetics and pharmacodynamics[4], and that could be also exploited for the production of new carbohydrate vaccines.

1.Introduction

1.1 Glycans, N-Glycosylation and glycosyltransferases

1.1.1 An overview of glycans

Glycans characterize every free-living cell in a multicellular organism. They cover the cell surface forming the so-called glycocalyx in eukaryotic cells, and they have an active role in several biological processes, including the infection by viruses, bacteria and parasites. Glycans also actively participate in the protein folding and quality control. The important role of glycans in human physiology is highlighted by the fact that defects on glycosylation processes result in pathologies classified as Congenital Disorders of Glycosylation (CDG) [5].

In Eukaryotic organisms, glycosylation is a post translational mechanism that originates in the ER-Golgi system and proceeds through the Golgi cisternae (also named the secretory pathway). Glycosylation of membrane proteins participates in cell-cell (or cell-pathogen) recognition and cell-matrix interaction, while glycosylation of secreted proteins may provide solubility, hydrophilicity and negative charge, thus reducing unwanted nonspecific intermolecular interactions in extracellular spaces and protecting against proteolysis. Glycans on secreted molecules may also act as decoys, binding pathogens that seek to recognize cell-surface glycans to initiate invasion[6].

In Bacteria, Archaea, and Fungi, glycans have critical structural roles in forming the cell wall and ensuring the correct osmolarity between the cytoplasm and the environment. Glycans surrounding bacteria could also have a role in defence against bacteriophages or antibiotics generated by other microorganisms in the environment[6].

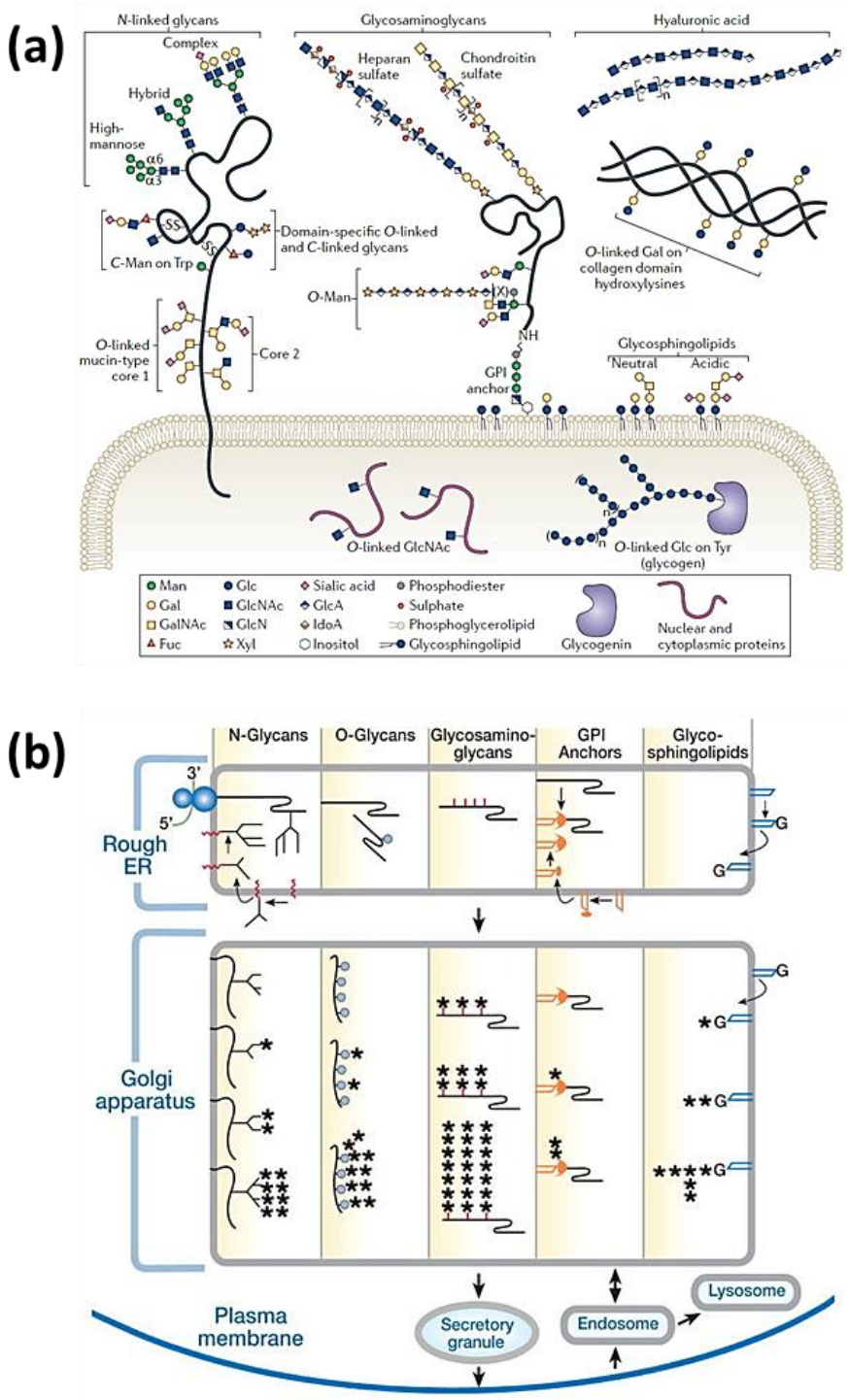


Figure 1. *Type of glycans and the Re-Golgi pathway*. **(a)** N-linked glycans, glycosaminoglycans and glycosphingolipids are located on cell surface proteins, hyaluronic acid is a component of external matrix, while O-linked glycans are usually situated on soluble proteins. (From Moremen et al.) **(b)** Initiation and maturation of the major types of eukaryotic glycoconjugates in relation to sub-cellular trafficking in the ER-Golgi–plasma membrane pathway. In addition to N-glycosylation, there are also O-glycosylation, glycosammination, the GPI anchor synthesis and the glycosphingolipids synthesis. (From Varki et al.[6])

Glycosylation is considered the most complicated post translational modification because of the number of reactions involved. Indeed, glycosylation molecular events include production of the donor substrates (either nucleotide-sugar or lipid linked precursors), its transfer to the nascent oligosaccharide chains and the trimming and the remodeling of the glycan, to obtain the final structure. Unlike other cell processes such as transcription or translation, glycosylation is non-templated, and thus all of these steps do not necessarily occur during every glycosylation event, amplifying the diversity of the final product[7].

As previously mentioned, in Eukarya through the ER-Golgi pathway proteins are modified by glycosyltransferases that transfer the lipid-linked precursor or an activated monosaccharide to a specific amino acid residue of the protein or to a growing glycan. Glycosidases catalyse the hydrolysis of glyosidic bonds to remove sugars from a glycan structure and finally, glycan-modifying enzymes transfer different groups from a specific donor to the glycan[7].

The **Figure 1** describes the different type of glycans and glycosylation processes among ER/Golgi system. **N-glycans** and **glycosylphosphatidylinositol** (GPI) anchors, a lipid anchor for many surface proteins, are initiated by the en-bloc transfer of a large preformed precursor glycan to a newly synthesized glycoprotein. **O-glycans** and **glycosaminoglycans** are initiated by the addition of a single monosaccharide, followed by sequential extension. The most common glycosphingolipids are initiated by the addition of glucose to ceramide on the outer face of the ER-Golgi compartments, and the glycan is then flipped into the lumen to be extended[6]. For more details on the other type of Glycosylation the reader is routed to the following references [6–8].

1.1.2 Protein N-glycosylation

N-glycosylation is the most highly studied form of protein glycosylation in eukaryotic organisms and it has been estimated that approximately half of all human proteins are glycoproteins, and most of them contain N-glycan structures[6]. All N-glycans share a common core sugar sequence (**Figure 2**), Manα1–6(Manα1–3)Manβ1–4GlcNAcβ1–4GlcNAcβ1-Asn-X-Ser/Thr (and less frequently of Asn-X-Cys and other non-standard sequons, where X can be any amino acid except for Pro[8]), and are classified into three types: (1) **oligomannose**, in which only mannose residues are attached to the core; (2) **complex**, in which "antennae" initiated by N-acetylglucosaminyltransferases (GlcNAcTs) are attached to the core; and (3) **hybrid**, in which only mannose residues are attached to the Manα1–6 arm of the core and one or two antennae are on the Manα1–3 arm[9].

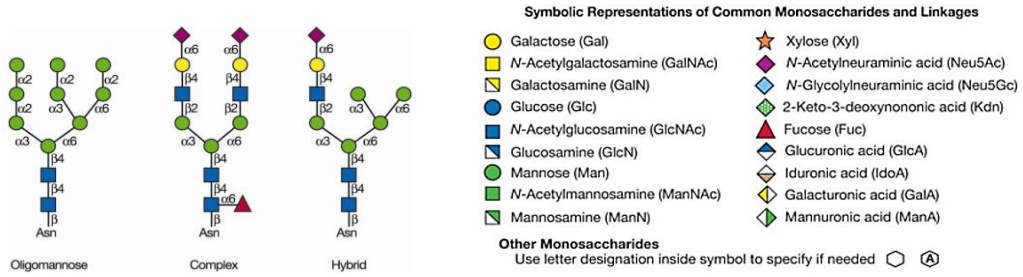


Figure 2. *N-glycans*. N-glycans added to protein at Asn-X-Ser/Thr sequons are of three general types in a mature glycoprotein: oligomannose, complex, and hybrid. Each N-glycan contains the common core Man3GlcNAc2Asn. (From Rini et al.[10])

As schematized in **Figure 3** and **Figure 4,** N-glycans are initially synthesized as a lipid-linked oligosaccharide (LLO) precursor (the dolichol phosphate or Dol-P in Eukaryotes, the undecaprenyl-pyrophosphate or Und-PP in bacteria), and then the glycans are transferred during translation from the precursor to a nascent polypeptide chain[8]. The biosynthesis begins on the cytoplasmic face of the ER membrane with the transfer of GlcNAc-P from UDP-GlcNAc to the Dol-P to generate dolichol pyrophosphate N-acetylglucosamine (Dol-P-P-GlcNAc)[9]. Fourteen sugars are then sequentially added to Dol-P, initially on the cytosolic face of ER and then, after flipping, on the ER luminal face. Finally, a multi-subunit

oligosaccharyltransferase (OST) on the luminal face of the ER membrane catalyses the "en bloc" transfer of the entire glycan to the Asn-X-Ser/Thr sequon on a protein that is being synthesized and translocated through the ER membrane (**Figure 3**)[8].

The protein-bound N-glycan is subsequently modified in the ER and Golgi by a complex series of reactions catalysed by membrane-bound glycosidases and glycosyltransferases. Many of these enzymes are finely sensitive to the physiological and biochemical conditions of the cell in which the glycoprotein is expressed. Indeed, activity of these enzymes will depend on the cell type in which the glycoprotein is expressed, may be regulated during development and differentiation and also it may be altered in disease[9,11].

It is important to note that whereas the presence of the Asn-X-Ser/Thr sequon is necessary for the formation of an N-glycan, transfer of the N-glycan to this sequon does not always occur, due to conformational state or other constraints during glycoprotein folding. Thus, the identity of "X" may reduce the efficiency of glycosylation, such as when "X" is acidic (aspartate or glutamate). In addition, when Asn-X-Ser/Thr sequons are present in an amino acid sequence, they are not identified categorically as N-glycan sites, but are referred to just a *potential* N-glycan sites[9].

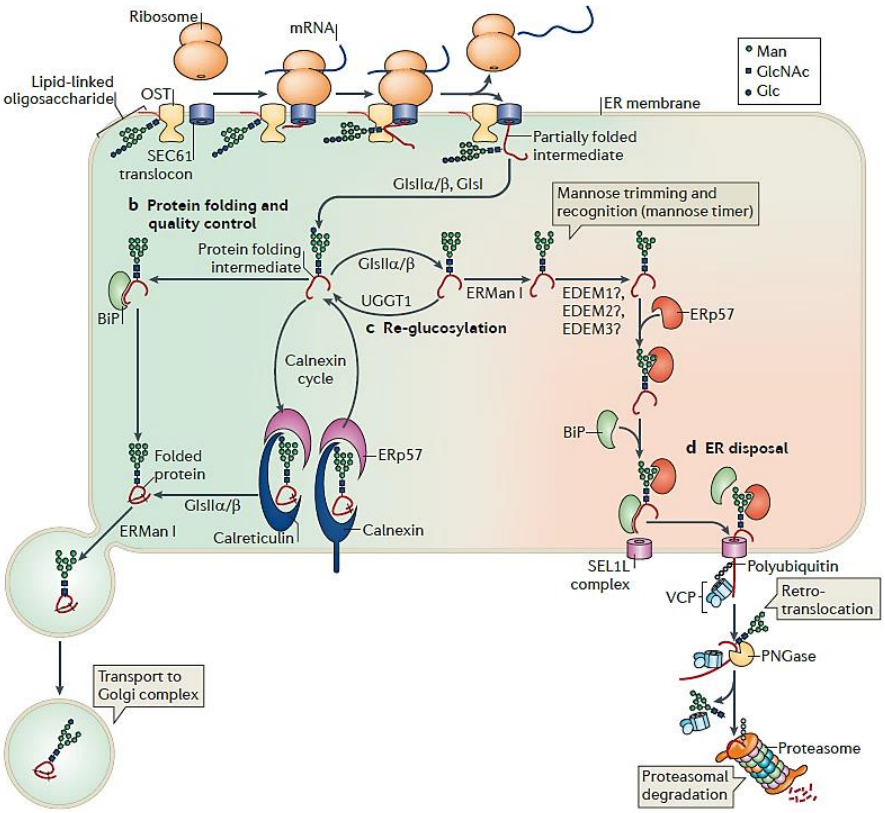


Figure 3. *Protein N-glycosylation and quality control of protein folding*. The STT3 subunit of the oligosaccaridiltransferase, transfers the glycan moiety on the nascent protein. Then, by the UGGT1 and the mannose trimmering process, the folding process involves lectin chaperons. If the protein could not fold, it is driven to the proteasomal degradation. On the contrary, the protein is translocated to the Golgi complex where glycosyltransferases will modify the glycans moiety (from Moremen et al.[8]).

As previously mentioned, the OST transfers the sugar moieties from the Dol-P to the nascent protein in eukaryotic cells. In particular, one OST subunit identified as STT3, contains the catalytic site of the enzyme. Bacterial glycosylation also relies on the transfer of glycans from a lipid-linked precursor to nascent polypeptide chains, but the enzyme responsible is a single polypeptide, PglB, with high sequence similarity to eukaryotic STT3[8].

The initial steps of the formation of the oligosaccharide precursor and its transfer on the nascent protein appear to be conserved among all eukaryotes. It is well established that the N-linked oligosaccharide has a key roles in protein quality process in the ER, via interactions with ER chaperones and lectins that recognize specific features of the trimmed glycan[9]. This process protects nascent polypeptides from hydrophobic aggregation and non-productive disulphide

bonding during their folding steps. Iterative cycles of glucose removal by glucosidase II and glucose re-addition by UDP-Glc:glycoprotein glucosyltransferase (UGGT1), followed by re-binding to the lectin chaperones calnexin and calreticulin, help to facilitate efficient folding of newly synthesized glycoproteins in the ER lumen[8]. In addition, to recruit chaperones during protein folding, glycan structures define the ER residence time for the downstream quality control of newly synthesized glycoproteins. ER-associated protein degradation (ERAD) is the process by which misfolded or unassembled proteins are destroyed in eukaryotic cells[8]. Glycoproteins with slow folding kinetics due to mutation or incomplete oligomeric assembly of subunits are targeted for degradation by the activity of mannose trimming enzymes, which is the basis of the 'mannose timer' model of ER quality control. Glycan trimming is followed by recognition of the trimmed structures by components of a multiprotein complex at the ER membrane that facilitates 'retro-translocation' and subsequent proteasomal degradation in a process termed as ER-associated degradation (ERAD)[8] (**Figure 3**). This complex process involves recognition of degradation signals, dislocation of proteins across the ER membrane and degradation by the ubiquitin-proteasome system in the cytoplasm[8].

Further processing of the N-linked glycan occurs in the downstream Golgi cysternae. Biosynthesis of hybrid and complex N-glycans is initiated in the *medial*-Golgi by the action of an N-acetylglucosaminyltransferase called GlcNAcT-I, which adds an *N*-acetyl-glucosamine residue to C-2 of the mannose α1–3 in the core of $Man_5GlcNAc_2$. Once this step has occurred, the majority of the N-glycans are trimmed by α-mannosidase II, another resident enzyme of the *medial*-Golgi, which removes the terminal α1-3Man and α1-6Man residues from $GlcNAcMan_5GlcNAc_2$ to form $GlcNAcMan_3GlcNAc_2$. It is important to note that α-mannosidase II cannot trim the $Man_5GlcNAc_2$ intermediate unless it is first acted on by GlcNAcT-I[9]. Once the two mannose residues are removed, a second N-acetylglucosamine is added to C-2 of the mannose α1–6 in the core by the action of GlcNAcT-II to yield the precursor for all biantennary, complex N-glycans.

Hybrid N-glycans (**Figure 3**) are formed if the $GlcNAcMan_5GlcNAc_2$ glycan is not acted on by α-mannosidase II, leaving the peripheral α1–3Man and α1–6Man

residues intact and unmodified in the mature glycoprotein. Incomplete action of α-mannosidase II can result in $GlcNAcMan_4GlcNAc_2$ hybrids. Another Golgi mannosidase, discovered in mutant mice lacking functional α-mannosidase II, is termed α-mannosidase IIX and acts on the $GlcNAcMan_5GlcNAc_2$ generated by GlcNAcT-1. Inactivation of both α-mannosidase II and α-mannosidase IIX in the mouse leads to embryos lacking all complex N-glycans[9]. Complex and hybrid N-glycans may carry a "bisecting" *N*-acetylglucosamine residue that is attached to the β-mannose of the core by GlcNAcT-III[9] (**Figure 3**).

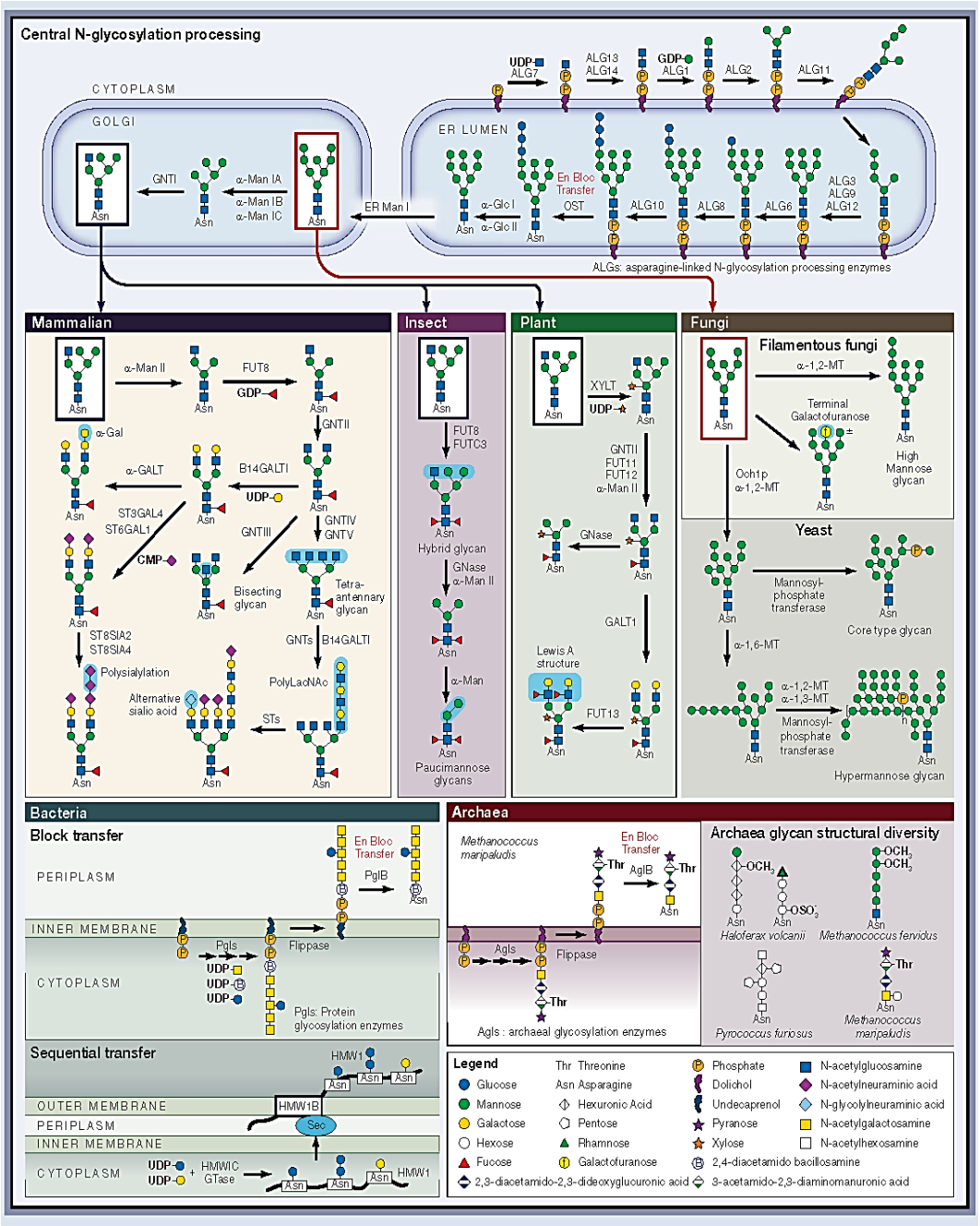


Figure 4. *The N-glycosylation process among kingdoms.* N-linked glycosylation in most eukaryotes follows a similar initial processing pathway, that begins with the generation of a lipidlinked oligosaccharide (LLO) by multiple asparagine-linked N-glycosylation processing enzymes (ALG). The oligosaccharide is then transferred "en bloc" on the polypeptide backbone by oligosaccharyltransferase (OST). Proteins are modified at Asn residues containing the N-X-S/T sequence. Processing then diverges significantly between evolutionarily distant species in the Golgi apparatus (From Cheng-Yu Chung et al. [12]).

Further sugar additions (**Figure 4**), mostly occurring in the *trans*-Golgi, convert the limited repertoire of hybrid and branched N-glycans into an extensive array of mature, complex N-glycans. For convenience, this part of the biosynthetic process can be divided into three components: **(1)** sugar additions to the core, **(2)** elongation of branching *N*-acetylglucosamine residues by sugar additions, and **(3)** "capping" or "decoration" of elongated branches. This steps are fully described here by Stanely et al. [9].

N-glycosylation is not restricted to Eukaryotes, but recent evidences have revealed complex mechanism of N-linked glycosylation also in Bacteria and Archaea, indicating that this type of modification is universal in all kingdom (**Figure 4**). In *Campylobacter jejuni* , the bacterium in which the N-glycosylation is deeply characterized, an heptasaccharide, the previously mentioned Und-PP, is built on a lipid-linked precursor on the cytoplasmic side of the inner membrane [13]. The resulting lipid-linked oligosaccharide (LLO) is then translocated across the inner membrane into the periplasmic space by an ATP-dependent flippase (PglK) and transferred to Asn residues in target proteins by a bacterial oligosaccharyltransferase (OTase), PglB. The bacterial system requires an extended N-glycosylation consensus sequence: Asp/Glu-X1 -Asn-X2 -Ser/Thr, where X1 and X2 represent any amino acid except Pro[13]. Finally, in Bacteria and Archea the monosaccharide bound to the Asn side chain is not GlcNAc, but acetyl-bacillosamine and other modified aminosugars are used.

Glycosyltransferases

In Eukaryotes most glycosyltransferases have been found to be localized in the ER-Golgi apparatus, but there are evidences of some cytosolic forms, such as the soluble OGT, a protein GlcNAc-transferase responsible for synthesizing the O-linked GlcNAc of nuclear and cytoplasmic proteins. Other forms of soluble glycosyltransferases are actually derived from their membrane-associated forms that is cleaved near the transmembrane segment within the stem region (**Figure 5**). These proteolytic cleavage events release a catalytically active fragment of the glycosyltransferase that is exported to the extracellular space[6] .

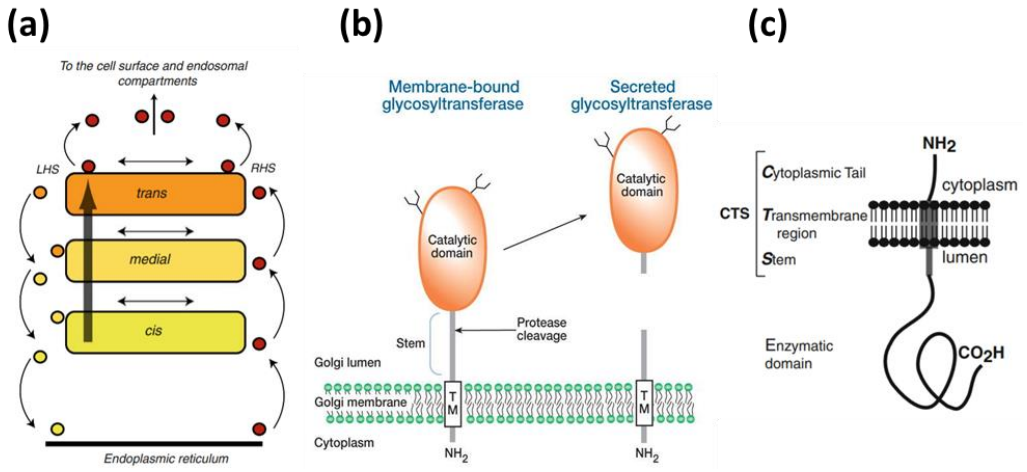


Figure 5. *Golgi organisation and typical transmembrane topology and proteolytic processing of Golgi glycosyltransferases.* **(a)** Golgi organelle is organised into three compartments: *cis, medial* and *trans.* The forward movement of vesicles starts from Endoplasmic Reticulum, until cell surface and endosomal compartments. Retrograde movement is the opposite (Modified from Tu et al.). **(b)** The soluble glycosyltransferases are cut near the transmembrane domain of the membrane –bound domain (Modified from Varki et al.). **(c)** Glycosyltransferases are composed by the enzymatic C-terminal domain that is exposed in the lumen, the transmembrane region and a short N-terminal domain that is exposed in the cytoplasm (Modified from Tu et al.[14]).

As the circulating enzymes do not have access to adequate concentrations of donor nucleotide sugars (primarily located inside cells), they should be functionally incapable of performing a transfer reaction in the extracellular spaces. For this reason, the biological significance of these soluble transferases therefore remains a mystery. Possibilities to consider include a lectin-like activity recognizing their acceptor substrates and/or a role in scavenging small amounts of

circulating sugar nucleotides that might otherwise be available to certain microbes, as it was demonstrated for gonococci[6].

As the **Figure 5** shows, typically the glycosyltransferases have a single transmembrane domain flanked by a short amino-terminal domain and a longer carboxy-terminal domain. This structure is characteristic of the so-called type II transmembrane proteins. The single amino-terminal membrane-spanning domain is a signal-anchor sequence, placing the short amino-terminal segment within the cytoplasm, while directing the larger carboxy-terminal domain to the other side of the biological membrane into which the signal anchor has been inserted. For plasma-membrane-associated type II proteins, the "other side" is the extracellular surface, but for glycosyltransferases, the "other side" is the lumen of the ER-Golgi pathway. Watching at this arrangement, it is evident that the larger carboxy-terminal domain contains the catalytic activity of the transferase, and the intralumenal location of this domain allows it to participate in the synthesis of the growing glycans, during transit of glycoproteins and glycolipids through the secretory pathway.

Generally, the Golgi glycosyltransferases localisation is mediated by different factors: the transmembrane domain, the luminally oriented noncatalytic region, interactions between catalytic domains, and the cytoplasmic tail. How the cytoplasmatic tail can drive Golgi-resident glycosyltransferases and glycosidases is not fully understood[14]. Biochemical and ultrastructural studies indicate that glycosyltransferases partially segregate into specifically distinct compartments within the secretory pathway. Generally, enzymes acting early in glycan biosynthetic pathways have been localized to *cis* and *medial* compartments of the Golgi, whereas enzymes acting later tend to co-localize in the *trans*-Golgi cisternae and the *trans*-Golgi network. In fact, the transmembrane domain of the Golgi glycosyltransferases possess a localisation signal and in many cases there are also major contributions from the luminal domain[15]. However, the mechanism is much more complicate, because it is also evident that different cell types from the same organism can show distinct localization patterns for the same enzyme [14]. Generally, Coat protein complex I (COPI)–coated vesicles are involved in the glycosyltransferases recycling[14]. The steady-state distribution of these enzymes is

maintained by a dynamic process that involves their retrieval from late Golgi cisterna (*Trans*-Golgi) to early cisterna (*Cis*-Golgi), as well as between the Golgi and the ER, then they transit back to the cisterna on which they function. However, Golgi-resident glycosyltransferases lack canonical COPI-binding motifs in their cytoplasmic tails, but there are other proteins, such as Sft1p, identified by Tu et al., that could bind the tails of these enzymes and could facilitate their incorporation into COPI-coated vesicles[14].

As glycosyltransferases can exist as monomers, homo-dimers, hetero-dimers, or hetero-oligomers, the glycosyltransferase associations appear to contribute to both the localization of these enzymes in the Golgi as well as to their enzymatic activity. At this point, homo-complexes, but more importantly hetero-complexes, play important role in the Golgi localisation. For example, the hetero-complexes allow glycosyltransferases to work in series as the product formed by one member of the complex serves as a substrate for another member[14]. As a demonstration of that, the sequential action of Mannosidase II (ManII) and N-acetylglucosaminyl transferase I (GlcNAcT1) is crucial to the synthesis of N-linked glycans in mammalian cells. In this way, the localization of one enzyme can influence the localization of the other [14] .

Actually, the formation of glycosyltransferase oligomers is not necessarily a prerequisite for proper localization, and therefore aggregation cannot account for the localization of all enzymes: the stem region and/or the transmembrane region and the physiochemical environment of particular cisternae may also influence the ability of glycosyltransferases to form oligomers as well as protein localization. This transition is highly dependent on the acidic Golgi microenvironment. Chloroquine, a pH gradient dissipating drug, markedly inhibited heteromer assembly in live cells[16]. Hassinen et al. demonstrated that after *de novo* synthesis, glycosyltransferase polypeptides form homodimers during their folding and/or before their transport to the Golgi. In the Golgi, the acidic luminal milieu favors the formation of heteromers among sequentially acting medial-Golgi or trans-Golgi enzymes at the expense of enzyme homomers. Nocodazole inhibits this process due to its ability to block anterograde transport and coalescence of

transport vesicles into a compact Golgi structure. Cloroquine was used to show that this transition is a pH-dependent process [16].

Glycosyltransferases structure and enzymatic mechanism

Glycosyltransferases comprise a large family of enzymes that share common features. The cloning and sequencing of more than 500 genomes has now shown that glycosyltransferases are a diffused enzyme type, representing 1–2% of the genes. More than 30,000 glycosyltransferase sequences are known across all kingdoms, and they comprise approximately 90 glycosyltransferase families defined by primary sequence analysis[10,17,18]

Despite the large number of sequence families that have been defined, structural analysis has shown that glycosyltransferases possess a limited number of fold types. To date, structures for members of 29 of the 90 families have been determined by X-ray crystallography, defining the GT-A or GT-B folds for the so-called Leloir enzymes[10].

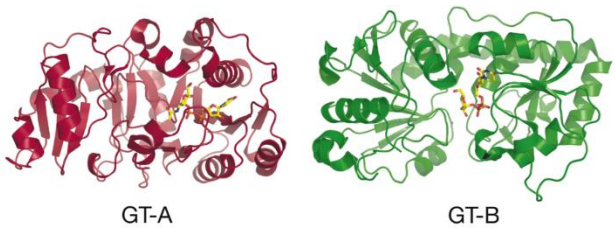


Figure 6. *Glycosyltransferases topological domains.* The GT-A-fold is a single catalytic domain arranged into two closely β/α/β Rossman domains. The GT-B-fold enzymes possess two distinct domains separated by a cleft that binds the acceptor in the middle. (From Rini et al.[10]).

GT-A-fold enzymes are composed by a single catalytic domain arranged into two closely β/α/β Rossman domains[19], which are found in proteins that bind nucleotides and are responsible for interaction with the nucleotide sugar donor substrate[10,19]. The GT-A enzymes have been found to possess a DXD motif, a short conserved motif that binds sugars, and are metal-ion-dependent glycosyltransferases[10].

GT-B-fold enzymes possess two distinct domains separated by a cleft that binds the acceptor. The carboxy-terminal domain is primarily responsible for binding the nucleotide sugar donor substrate, but both domains possess elements similar to those of the Rossman fold. The GT-B glycosyltransferases are often metal-ion independent and do not possess a DXD motif [10,19].

Leloir type glycosyltransferases transfer a nucleotide-activated sugar to the nascent glycan moiety. The activated donor sugar substrate contains a (substituted) phosphate leaving group; in fact, they are nucleoside diphosphate or monophosphate sugars (e.g., UDP-galactose, GDP-mannose, CMP-sialic acid). Finally, glycosyltransferases use as acceptor substrates oligosaccharides, monosaccharides, polypeptides, lipids, small organic molecules, and even DNA[19,10].

As previously described, these enzymes act sequentially, so that the product of one enzyme yields the acceptor substrate for the subsequent action of another. The end result is a linear and/or branched polymer composed of monosaccharides linked to one another[10].

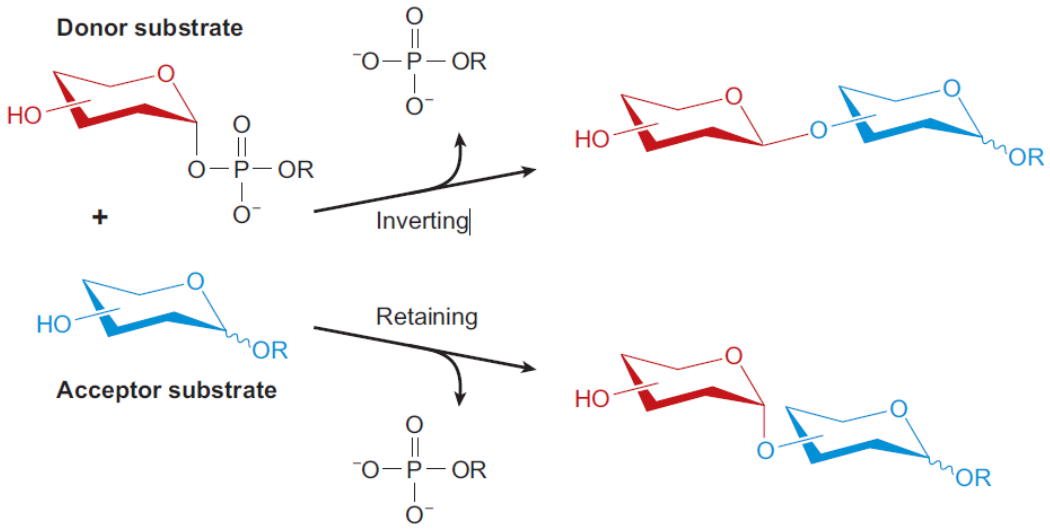


Figure 7 . *Glycosyltransferases catalyse glycosyl group transfer with either inversion or retention of the anomeric stereochemistry, with respect to the donor sugar.* In the inversion stereochemistry, the enzyme uses an S_N2 reaction, while the retention configuration follows the so-called S_Ni mechanism (From Rini et al. [10]).

As shown in **Figure 7**, glycosyltransferases catalyse their reactions with either inversion or retention of stereochemistry at the anomeric carbon atom of the donor substrate. In the **inversion** stereochemistry, the enzyme uses an S_N2 (substitution nucleophilic bimolecular) reaction mechanism where the acceptor performs a nucleophilic attack at carbon C-1 of the sugar donor. Typically, enzymes of this type possess an aspartic acid or glutamic acid residue whose side chain serves to partially deprotonate the incoming acceptor hydroxyl group, rendering it a better nucleophile. In addition, these enzymes promote catalysis by features that help to promote leaving-group departure. In the GT-A enzymes, a metal ion is bound by the DXD motif and it is typically positioned in the way to interact with the diphosphate moiety. The positively charged metal ion serves to electrostatically stabilize the additional negative charge that develops on the UDP leaving group during bond breakage[10].

The **retention** reaction follows the so-called S_Ni mechanism. In this case, the incoming nucleophile attacks from the same side as the leaving group. It means that the leaving group departure and nucleophilic attack occur in a concerted but asynchronous manner on the same face of the glycoside[10].

Many glycosyltransferases have been shown to possess a Bi-Bi sequential kinetic mechanism in which the donor substrate binds before the acceptor substrate, and the glycosylated acceptor is released before the nucleoside monosphosphate or disphosphate, depending on the reaction. To better explain this structural model, the active site represents a deep pocket, with the nucleotide sugar substrate at the bottom and the acceptor substrate stacked on top. If the acceptor substrate binds first, it would sterically preclude donor substrate binding. Necessarily, release of the glycosylated product must precede release of the nucleoside phosphate. Although largely consistent with such a model, the X-ray crystal structures of glycosyltransferase-substrate complexes also shows that substrate-dependent ordering of flexible loops is a feature common to glycosyltransferases. Typically, donor substrate binding orders a loop(s) that in turn facilitates acceptor substrate binding[10].

The Biotechnological interest in Glycosyltransferases

The study of glycosyltransferases is of particular biotechnological interest because of the functional contribution of carbohydrates as biologically active natural products. Actually, there are several glycosides produced by microorganisms and plants which are used as drugs in the treatment of different diseases[20].

Moreover, as the therapeutically relevant natural products are glycosylated, and because of the sugar residues attached to such natural products by glycosyltransferases are typically indispensable for biological activity, the exact identity and pattern of glycosyl moieties can influence pharmacology/pharmacokinetics, invoke biological specificity at the molecular/tissue/organism level, and even define the precise mechanism of action. This fact, coupled with the importance of natural products in drug development, has spurred the development of both chemical and enzymatic methods for glycosylating natural products[21].

Functional characterisation of each glycosyltransferase is required to explore the potential of these enzymes for the derivatisation of glycosylated natural products and with the advent of molecular tools and recombinant methods it is now possible to engineer novel natural product derivatives.

In the biosynthesis of bioactive natural glycosides, glycosyltransferases are often involved in the last-step modification of an aglycon, leading to the corresponding ultimate bioactive molecules. However, the substrate specificity of glycosyltransferases provides a critical issue in natural product diversification, and scientists have started recently to broaden the specificity by genetic engineering[34].

Recently, glycosyltransferases with broad tolerance, especially towards the sugar donor, have been identified and characterised making them valuable as tools, for example, in antibiotic remodeling[20].

To date, carbohydrate vaccines are used in vaccines industry. Usually, they are directly purified from the target microorganism and the presence of impurities (such as cell-wall polysaccharides) are coisolated, with the risk of hyperresponsivity and other side effects [22,23]. The isolated polysaccharides are structurally heterogeneous, and they require multiple purification and quality

control steps before that the antigen can be formulated in a vaccine. Efforts to improve the efficacy and safety of these vaccines are important to achieve comprehensive vaccination and eradication of the respective pathogens.

Synthetic oligosaccharides based on the repeating units can be an attractive option to furnish vaccines free of contaminants that have predictable clinical outcomes. An interesting approach, that will avoid the copurification of impurities from microorganisms or the difficulties linked to the chemical synthesis, could be the *in vitro* production of oligosaccharides by using recombinant glycosyltransferases. However, most glycosyltransferases are membrane bound and difficult to produce as soluble enzymes in *E. coli.* The viral glycosyltransferases from giant viruses studied in this work are soluble enzymes and can be produced in large amount and this peculiarity could be exploited in the production of complex carbohydrates.

1.2 Nucleo cytoplasmic large DNA viruses

"Giant viruses sound like something from a scifi flick. But they're real and not as scary as you think".
Jim Van Etten

1.2.1 Evolution theories and families of Nucleo Cytoplasmic Large DNA Viruses

Viruses were always considered as not "live" organisms, but such as entities composed by a capsid (and an envelope in most cases) that contains the genetic material (DNA or RNA). As suggested by Patrick Forterre[24], viruses are viewed also as "side-products of cellular evolution". But now, in the recent 20 years, viruses are the center of many debates on the early evolution of life on our planet and the question "are viruses alive?" is at the center of this debate. For many years the answer to this question was negative, considering viruses as "escaped genes" from cellular organisms[24].

Viruses infect organisms in all superkindoms of life (Bacteria, Archaea and Eukaryotes) and replicate in all cell type[25]. Their extreme diversity suggests that they must have had a multiple evolutionary origin. According to Iyer et al [26], there are two major group of theories for virus evolution: the first one places viruses in the earliest phases of life's evolution and associate them with the primitive precursors of cellular systems. The second group of theories consider viruses as secondary derivatives of cellular systems that underwent to a drastic degeneration as a consequence of extreme parasitism. Moreover, viruses are seen as "break away" elements from cellular genomes that survived as minimal parasitic replicons[26].

The first decade of viral comparative genomics unifies viruses on the basis of the evolutionarily conserved proteins of their replicon apparatus [26]. Here we find the division between **retroviruses**, that use retrotransposons and a reverse transcriptase as their principal replicon polymerase, and **DNA viruses**, with their related plasmids and transposons, that use replication endonucleases [26]. Inside these two big classifications, there are several monophyletic groups with common ancestors, as the **Nucleo Cytoplasmic DNA Viruses (NCLDVs)**[26]. However, the

higher order relationship between various groups of large eukaryotic DNA viruses remain uncertain.

NCLDVs are a group of DNA viruses infecting diverse hosts and that share common ancestry. Eight families that infects eukaryotes (algae, animals and protista) have been recognized, represented by *Poxoviridae* that infect vertebrates and invertebrates*, Asfaviridae* that infect vertebrates*, Iridoviridae* and *Phycodnaviridae* that are aquatic viruses, *Ascoviridae, Marseilleviridae, Pithoviridae* and *Mimiviridae*[25,27].

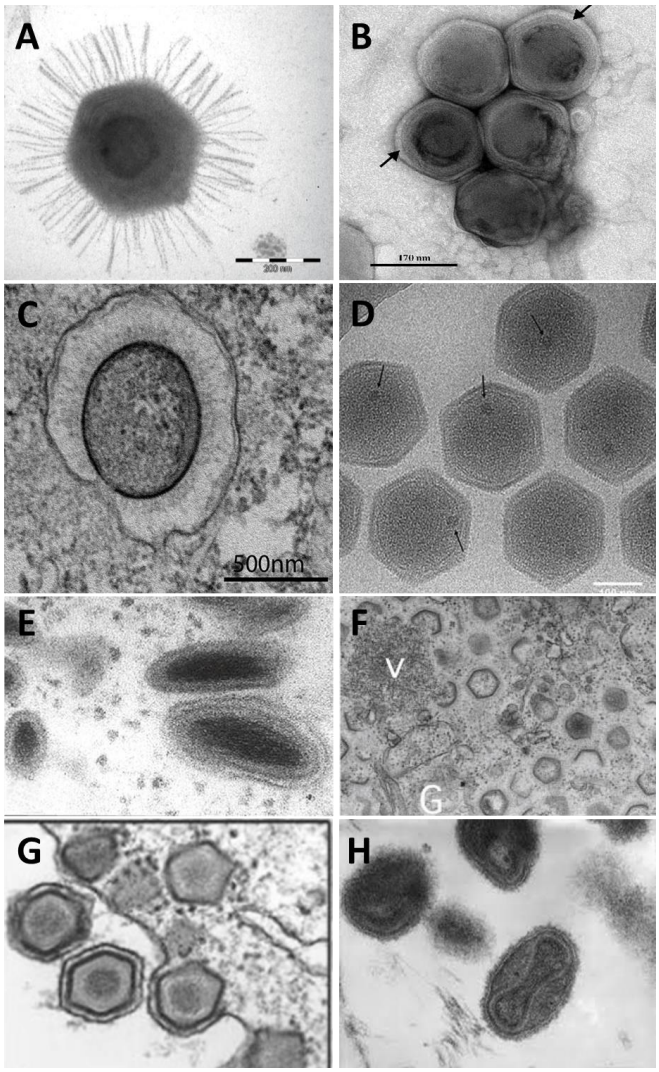


Figure 8. *The eight families of NCLDVs classified so far.* **(A)** Mimivirus (Modified from Wilson et al[28].) **(B)** Phycodnavirus (Modified from Wilson et al[28].) **(C)** Pithovirus (Modified from Abergel et al[29].) **(D)** Marseillevirus (Modified from Okamoto et al[30].) **(E)** Ascovirus (Modified from Wilson et al[28].) **(F)** Iridovirus (Modified from Wilson et al[28]) **(G)** Asfavirus (Modified from Wilson et al[28].) **(H)** Poxovirus (Modified from Jha et al.[31]).

Mimivirus (**Figure 8**, panel (A)) is the first giant virus that was discovered in 1992 by Timothy Robotham, a microbiologist at Leeds Public Health Laboratory[1]. In 2003 Mimivirus was isolated and sequenced after the observation, thanks to electron microscopy techniques, that it was not a "clamidia-like" bacteria, but a giant virus with 1.1 megabases of genome and 0.7 μM of dimensions.[29] Mimivirus is characterized by an icosahedral capsid of about 400 nm of dimensions, surrounded by 150 nm of thick fibril layer with a composition similar to peptidoglycan, except for a five-pronged star structure in a vertex defined as "stargate"[29]. This structure is used by the virus to fuse with Acanthamoeba membrane and inject the viral nucloide inside the host cytoplasm[29].

Mimivirus genome analysis was not able to places it in the known domains of life, suggesting an independent evolution from a different ancestor that belongs to a new undefined tree[32]. The new, fourth, domain of life is strongly debated on both technical and biological aspects, evolving into a quasi-philosophical debate on the nature of viruses, as previously mentioned in this chapter: "are viruses alive?" and "could viruses, in principle, belong to a tree of life?"[24,27]. Starting from Mimivirus, a lot of NCLDVs are then discovered, isolated and genome sequenced, as briefly described in this chapter. But recently, the interest in giant viruses is grown also thanks to the discovery of viriophages, small dsDNA viruses classified as *Lavidaviridae* that parasitize Mimivirus and reproduce in its viral factories and can block its growth, partially defending the Amoeba host[27,33]. This finding, and other characteristic such as genome size and viral particle dimensions, define Mimivirus and all the NCLDVs as viruses with cell-like properties.

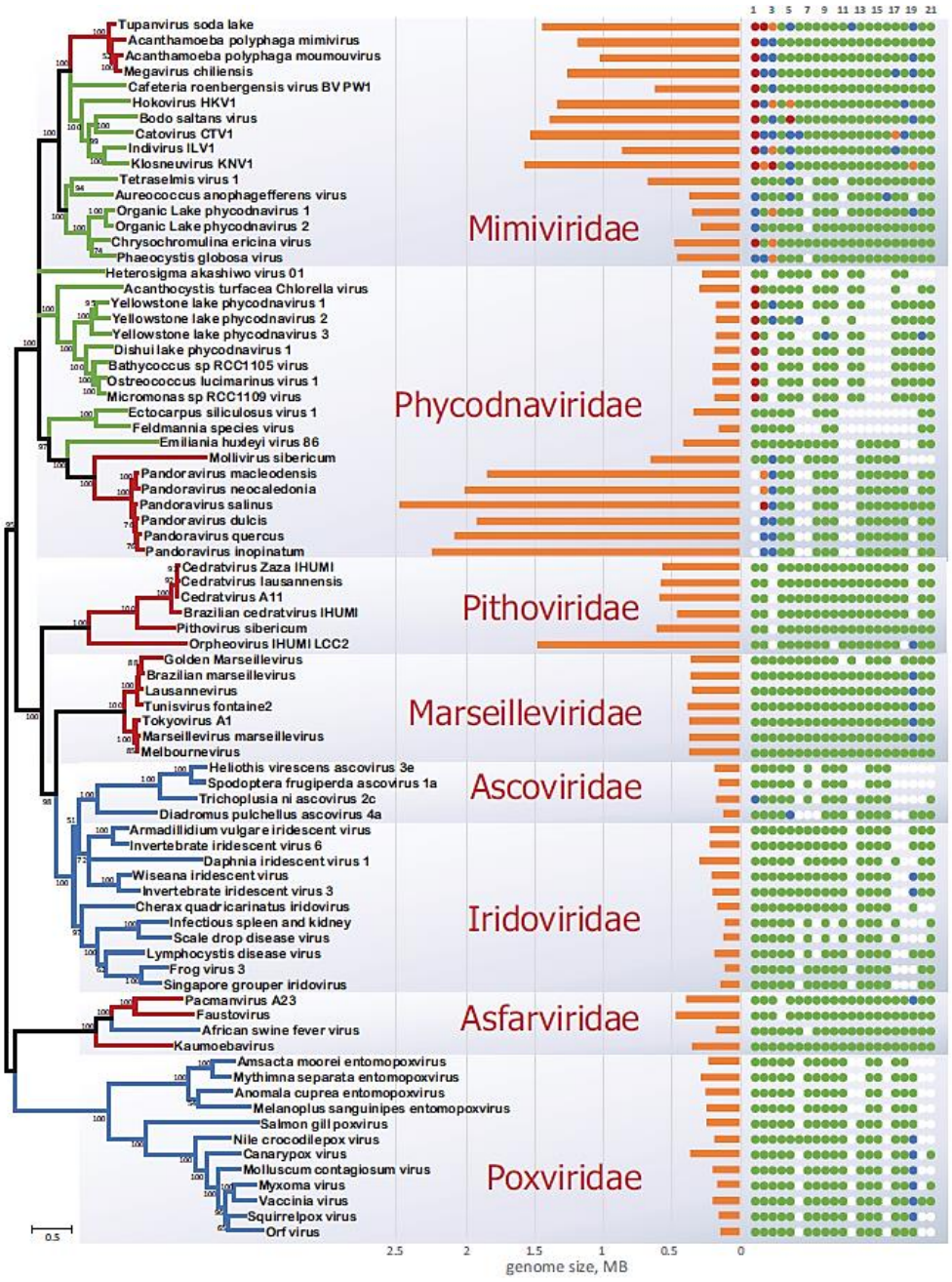


Figure 9. *The family relationships among NCLDVs members are represented by the phylogenetic reconstruction of universally conserved NCLDVs proteins: DNA polymerase, major capsid protein, packaging ATPase, A-18 like helicase, Poxovirus late transcription factor VLTF3* . The branch colour indicates confirmed or likely hosts: red Amoebozoa; green other protists; blue Metazoa. The tree reconstruction is a phylogeny obtained with an updated, representative set of NCLDVs according to the new discoveries. It consists in three big branches: families of *Mimiviridae* and *Phycodnaviridae*, families of *Pithoviridae*, *Marseilleviridae*, *Iridoviridae* and *Ascoviridae*, and finally families of *Poxoviridae* and *Asfaviridae*. (From Koonin et al.[27]).

Poxviruses (**Figure 8,** panel (H)) infect animals and replicate entirely in the cytoplasm. [34] They possess a unique icosahedral capsid characterised by brick-shaped virions. They are classified in a clade that contains also **Asfavirus** [27] (**Figure 8** panel (G))**.** As described by Koonin et al[27], the switch from protists to animal hosts seem to be appeared twice in this branch. Finally, Poxvirus contains genome that spaces from 130 to 360 kb [27] (Table1).

Asfavirus are in the same clade of Poxoviruses with share common characteristics such as the icosahedral envelope and the cytoplasmic replication. They infect amoebae and animals and possess dsDNA of 170-470kb[27,35] (Table1).

Iridovirus (**Figure 8,** panel (F) are in the same branch of **Ascovirus** (**Figure 8,** panel (E)) **, Marseillevirus** (**Figure 8,** panel (D)) and **Pithovirus** families (**Figure 8,** panel (C), and **Figure 9**). [32] Based on phylogenic studies, Iridoviruses and Asfaviruses evolved from Phycodnaviruses, and the Ascoviruses evolved from Iridoviruses[36]. Iridoviruses infect insect and cold-blooded vertebrates and have genome of about 250kb with icosahedral capsid. They can replicate in both host nucleus and cytoplasm[32] (Table1).

Ascoviruses produce virions that are reniform or bacilliform and cause fatal disease in their insect hosts[27,36]. They have a 150kb DNA and replicate in nucleus and cytoplasm [33] (**Figure 8**, panel (E), Table1).

Pithoviruses have an amphora-shaped form and infect probably protist. They were discovered using the same protocol for coculturing acanthamoeba and Pandoravirus, by which was confused thanks to the same viral dimensions and shape[29]. Pithoviruses have the largest virions among all known virions with genome that can reach more than 1400kb [27] (**Figure 8**, panel (C), Table 1).

Marseillevirus is one of the last NCLDVs discovered and it is recently established that is a member of the amoebal NCLDVs. [30] The virions shows ichosaedral capsid that can vary from 190 to 250 nm, but the anatomic and genic carachteristics place them in a unique family [30] . Finally, the Marseillevirus genome span around 350kb and replicate in host nucleus and cytoplasm (Table 1)[27].

Phycodnaviruses (**Figure 8,** panel (B)) that infect the green algae *Chlorella variabilis* [28], hold genome size until 400 kb and replicate in both nucleus and

cytoplasm. Phycodnavirus, and more specifically PBCV-1, is the main object of this thesis and will be discuss largely in the next pages (Table 1).

Recently, some new viruses have been discovered. More precisely, they are Faustovirus, Mollivirus and Pandoravirus (**Figure 10**).

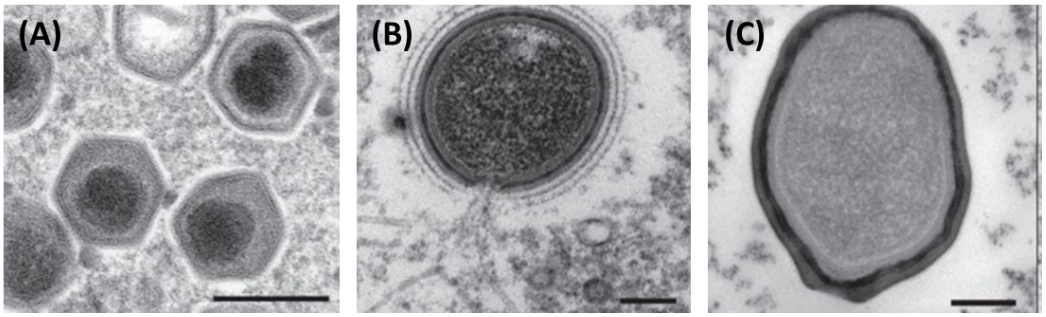


Figure 10. *New NCLDVs.* **(A)** Faustovirus. **(B)** Mollivirus. **(C)** Pandoravirus. Scale bar: 200 nm. (Modified from Fisher et al[37].)

Pandoravirions (**Figure 10**, panel (C)), recently classified id the *Pithovirus* family, (**Figure 8**) have a novel type of morphology, consisting of 1 mm by 0.5 mm ovoid particles with a single apical pore, needed for genome delivery. Comparative genomic analyses confirmed that Pandoraviruses are unrelated to Mimiviruses, but they share a few genes with algae-infecting phycodnaviruses[37]. The main peculiarity of Pandoravirus is a compartment between two layers, resulting in a 70 nm tegument like envelope of three layers [29]. Finally, Pandoravirus DNA consists of about 3 Mbp. [29]

From Siberian permafrost (dated to be 30 000 years old), two novel types of giant viruses were isolated on Acanthamoeba hosts. *Pithovirus sibericum* (**Figure 8** panel (C)) displays an elongated particle with a length of 1.5 mm and a diameter of 0.5 mm, which resembles pandoravirions, but are structurally unrelated. Phylogenetically, this virus shows a remote relatedness to irido- and ascoviruses.

The second novel type of giant virus that was isolated from permafrost soil is **Mollivirus sibericum** (**Figure 10** panel (B)), not assigned to any family yet, with previously unseen morphology. They are spherical, 0.6 mm in diameter, and covered with a 'hairy' tegument that can be from two to four [29,37]. Mollivirus apex aperture consists then in a funnel of about 200 nm in diameter[29].

Another described addition to the NCLDV group is **Faustovirus** (**Figure 10** panel (A))**,** which was isolated on *Vermamoeba vermiformis*, an amoeba that is commonly found in human environments. Although the 'Faustovirus' capsids (about 200nm) have the typical icosahedral symmetry also found in the amoeba infecting mimiviruses or marseilleviruses, their 466 kb genome encodes several proteins with phylogenetic affinity to asfarviruses (African swine fever virus).

Origin and evolution of NCLDVs is still unclear. Surely, there is a group of genes that, for the majority, are of eukaryotic origins and just a minority of them are from a bacterial and archaeal descendent. For this reason, it is supposed that they have coevoluted during eukaryogenesis. Lineage- specific gene loss and gain within NCLDVs families has led to the highly diverse properties of present-day viruses[25]. In the beginning of 2000s, the evolution of NCLDVs was supposed to come from a single ancestor, thanks to the observation of a core gene set of about 50 genes [25,26]. More recently, it is suggested by Koonin et al. that giant viruses have encountered a dynamic evolution, and not from a single anchestor[27]. According to the first vision of NCLDVs evolution, the first 50 genes might be retained or discarded in a distinctive evolutionary traceable pattern that could be mapped to the genome of the common ancestor of this class of eukaryotic viruses[1,25,28]. Also, according to Claverie et al. as they supposed in the first decade of 2000, giant viruses must come from a cellular ancestor that could not be assigned to any of the three domains of life, but from a separated fourth one[38]. This finding came primarily from the observation that, mapping Mimivirus DNA, it is represented by a complex genome not characterised by accumulation of random DNA segments (like in bacteria) or particularly enriched with mobile elements, palindromic structures, or genes encoding the necessary enzymatic equipment (transposases, integrases...)[38]. Moreover, as usually happens for conventional viruses, there is no presence of lateral gene transfer in the so far studied Mimivirus genome (and later for the others)[38]. The concept of a fourth domain of life has been promoted starting from 2011 from different scientist, with a new concept of NCLDVs evolution, as will be described in the next pages[27].

Virus Family/Group	Host Range	Genome Size Range (kb)	Virion Architecture	Replication Site
"Extended Mimiviridae" Mimiviridae Proposed subfamily "Klosneuvirinae" OLPG group	Acanthamoeba and, probably, other amoebae; algae, heterokonts	280–1570	Icosahedral	Cytoplasm
Phycodnaviridae	Green algae; algal symbionts of paramecia and hydras; heterokonts; Haptophyta	180–400	Icosahedral	Nucleus and cytoplasm
"Pandoraviridae" Mollivirus sibericum Pandoraviruses	Amoebae	650–2470	Spherical; Amphora-shaped	Nucleus and cytoplasm
Pithoviridae	Unknown protists	460–1470	Amphora-shaped	Cytoplasm
Marseilleviridae	Acanthamoeba; probably, also algae	360–380	Icosahedral	Nucleus and cytoplasm
Asco-* and *Iridoviridae	Invertebrates and non-mammalian vertebrates	100–290		
Ascoviridae	Insects, mainly, Noctuids	120–190	Ovoid	Nucleus and cytoplasm
Iridoviridae	Insects, cold-blooded vertebrates	100–290	Icosahedral	Nucleus and cytoplasm
"Extended *Asfarviridae*" Asfarviridae Faustoviruses Pacmanvirus Kaemoebavirus	Amoebae, mammals	170–470	Icosahedral	Cytoplasm
Poxviridae	Animals: vertebrates, insects	130–360	Brick-shaped, icosahedral intermediate	Cytoplasm

Table 1. *NCLDVs families with unrelated groups.* The table shows the main NCLDVs characteristics in terms of host, genome size, virion architecture and replication site. (From Koonin et al[27].)

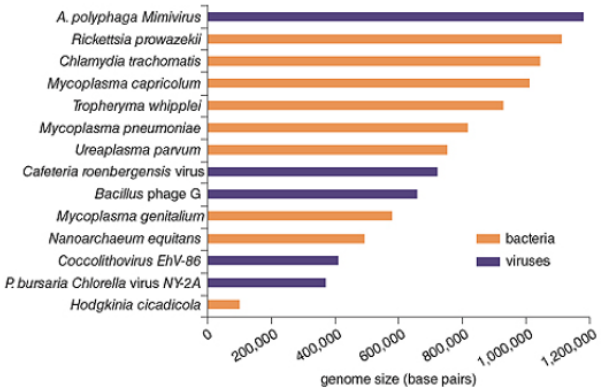


Figure 11. *Different viral genomes size comparison.* Mimivirus is a giant among giant viruses, with a diameter of 750 nanometers. It possesses a very big genome comparing to viral standards, of 1.2 million base pairs, coding for 1,018 genes. For comparison, the smallest free-living bacterium, Mycoplasma genitalium, is just 450 nanometers in diameter and possesses a genome half the size of that in mimivirus, coding just 482 proteins. The record tiniest cellular organism, Hodgkinia cicadicola, a parasite in cicadas that was described in 2009, has a genome of just 140,000 base pairs, coding a paltry 169 proteins (from Van Etten et al[1].)

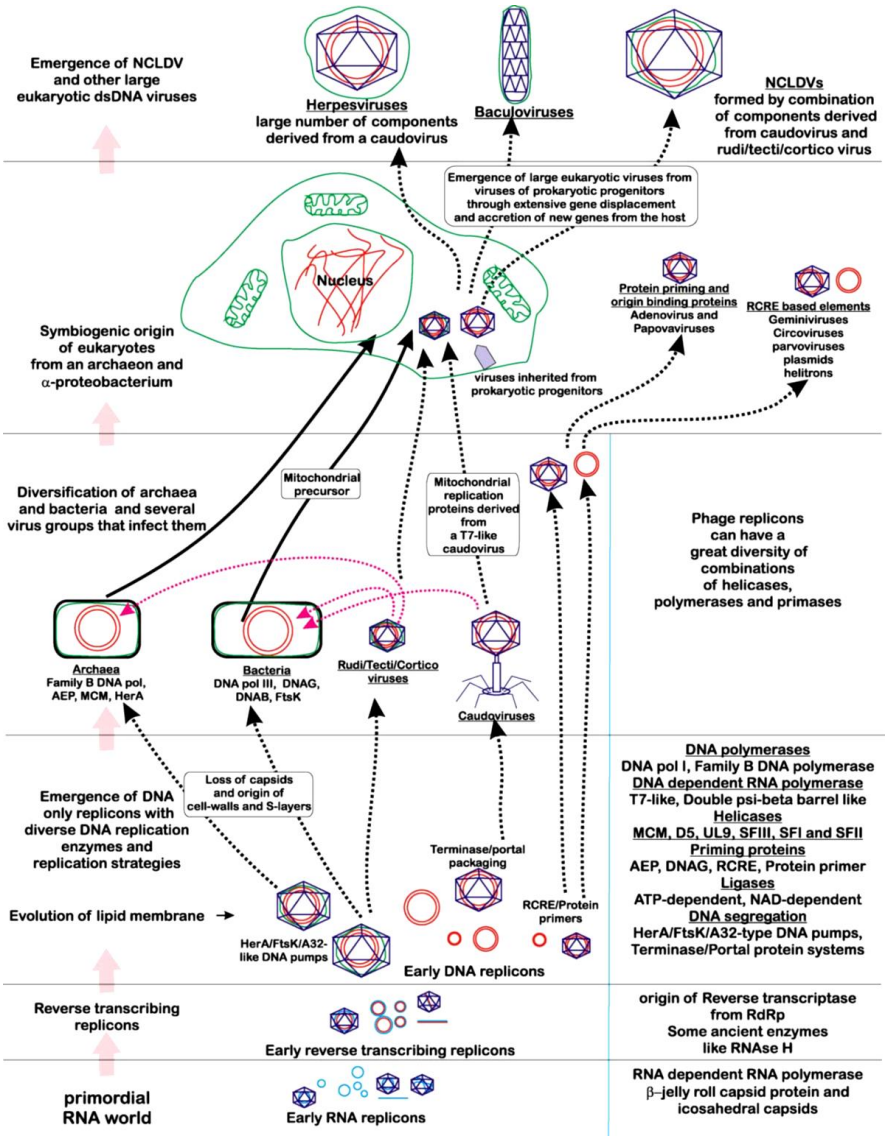


Figure 12. *A schematic representation of the probable scenario of evolution of DNA viruses and other DNA-based replicons proposed by Iyer et al. in the first 2000s.* On the left are pointed some of the major transitions in evolution. On the right are briefly described the major innovations entailed by each transition. DNA genomes are colored red, RNAgenomes are colored blue, and lipid membranes are colored green (from Iyer et al[26].)

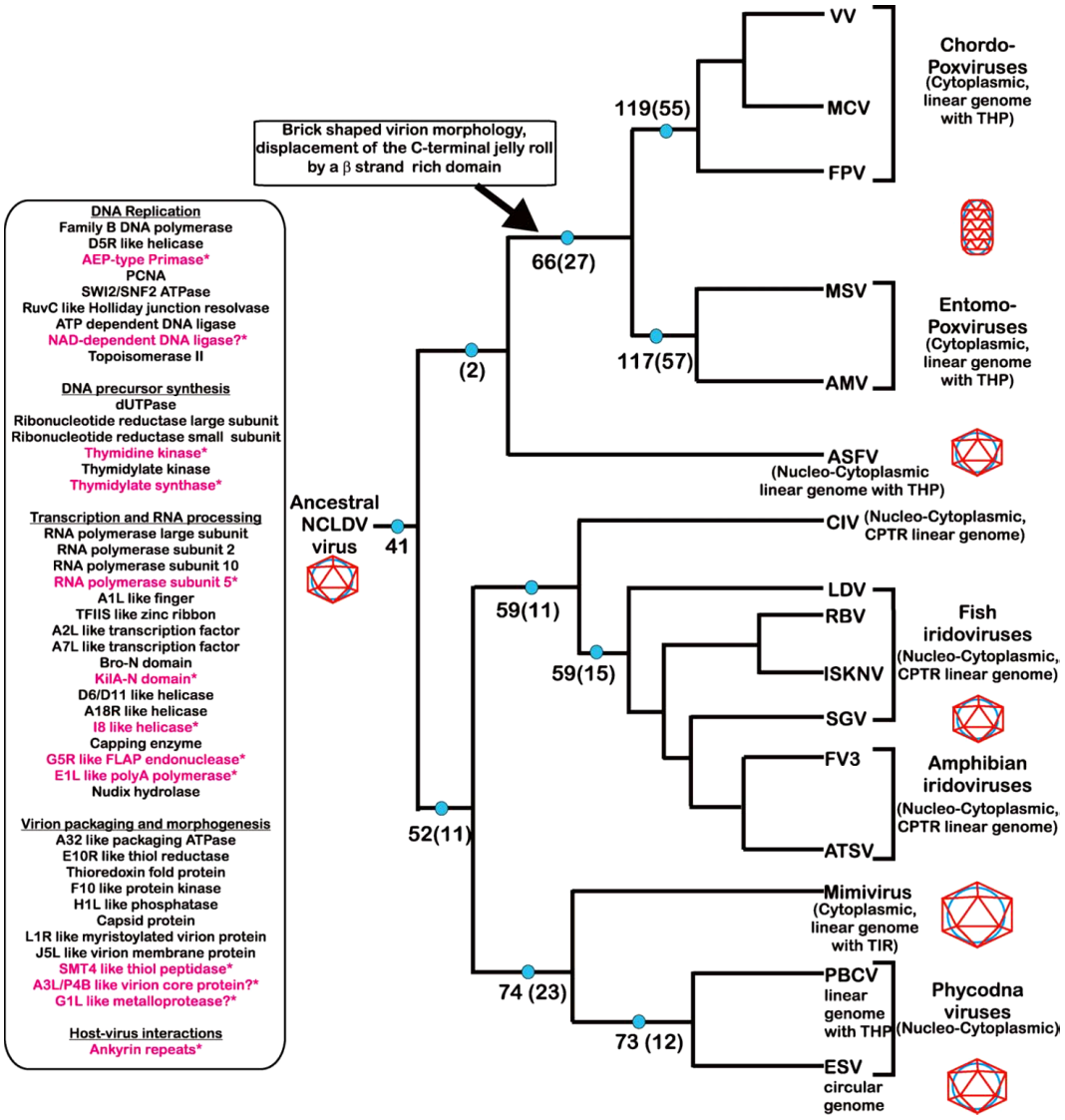


Figure 13. *Iyer et al propose also a phylogenetic tree of the NCLDVs, built on the basis of conserved gene set analysis*. The number of proteins reconstructed as being present in the ancestral core of a clade of viruses is shown next to the blue circles. Shown in brackets are the numbers of proteins that are unique to a particular clade. Proteins that are predicted to be part of the ancestral NCLDV genome are shown on the left. Protein names in red and marked with an asterisk represent members of the ancestral genes set that have not been identified in previous reconstruction of the author. The transcription factor A7L protein was formerly called the ASFV-B385R-like protein (abbreviations: THP, Terminal hairpin; TIR, Terminal inverted repeats; CPTR, circularly permuted terminally redundant. from Iyer et al[26]).

As previously mentioned, the new vision of NCLDVs evolution proposed by Koonin et al. [27] suggests a **dynamic evolution** instead of an evolution from a single ancestor, thanks also to the discovery of new species of giant viruses that allowed to amplify the NCLDVs genome analysis and knowledge. In the model proposed and well explained by bioinformatic analysis, the NCLDVs pangenome (the entire set of genes for all strains in a clade), is "heavily dominated by ORFans and genes that are conserved in small groups of viruses only", infact, the overlap between gene sets is demonstrated to be small (**Figure 14**) [27].

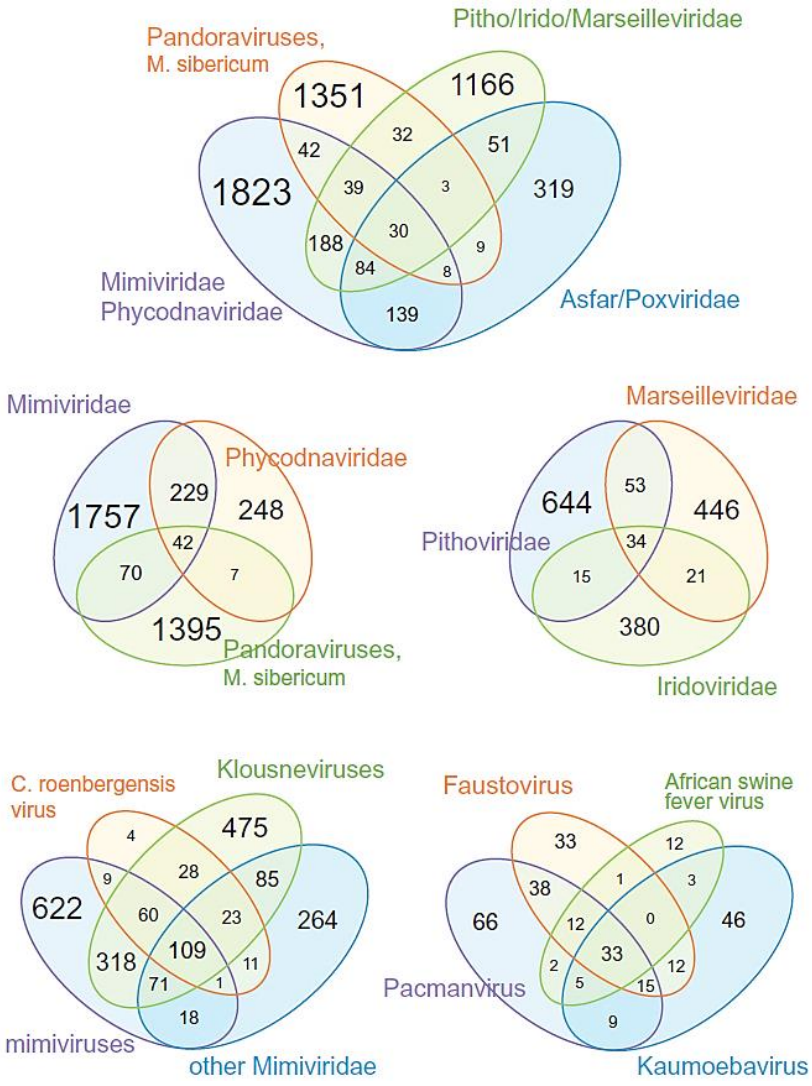


Figure 14. D*ifferently from how it was proposed from Iyer et al in 2001, that suggested a NCLDVs evolution from a single ancestor, later in 2006 it is proposed a multiple coevolution from at least three different organisms.* This finding came from the observation that there are few genes conserved in all member of each group. (From Koonin et al[27]).

To explain this dynamic evolution, Koonin et al. suppose that both gene gain and gene loss might be occurred (**Figure 15**)[27]. In prokaryotes, losses are more common than gains, differently from how it probably happened in NCLDVs evolution where gains are prevailing. This observation is confirmed by the likelihood reconstruction of genome evolution along the tree of the nearly universal genes in giant viruses[27]. As described above, and how it is possible to observe in **Figure 15**, in NCLDVs there is a lineage specific gains rather than losses of ancestral genes. For this reason coupled with the small overlapping of genes between NCLDVs genomes, it is supposed an evolution from at least three independent occasions, differently from the previous proposed view [25,26]. However, the origin and evolution of NCLDVs remains an open debate.

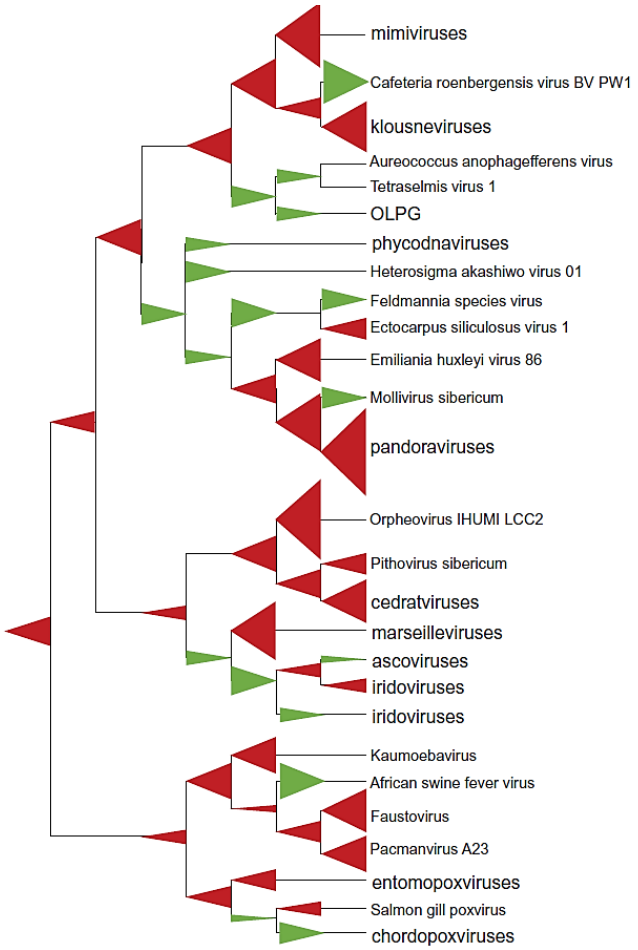


Figure 15. *Gene gain and gene loss in NCLDVs evolution.* Red triangles represent gene loss while green triangles gene gains. It is also possible to appreciate the three main branches of NCLDVs families (From Koonin et al[27].)

The large genome of NCLDVs comprises proteins with cell like proprieties that give a partial independence from the cell host. For instance, it is demonstrated that they possess the replication machinery, proteases, part of the translation system, enzymes involved in redox reactions and, of particular interest, the enzymes involved in glycosylation and modification of glycans [39]. This viewpoint is important comparing NCLDVs to common viruses, which despite the enzymes needed for the infection and, in some cases, DNA replication and translation, they are totally dependent from the host.

NCLDVs replicate exclusively in the host cytoplasm or start their life cycle in the host nucleus and complete it in the cytoplasm, as evidenced for *Mimivirus* and *Phycodnaviridae* [26,40,41]. In addition, they typically do not exhibit much dependence on the host replication or transcription systems (as usually viruses do) because irradiation of host nucleus does not inhibit replication of viruses, as Professor Van Etten demonstrated for *Paramecium bursaria Chlorella virus – 1*(PBCV -1) [42]. In line of that, the NCLDVs encode several conserved proteins that mediate most of the processes essential for viral reproduction.

As previously mentioned, NCLDVs infect animals, algae and Protista and their peculiarities are a high genome complexity and notable viral particle dimensions. As an example, PBCV– 1 displays a 300 kb DNA that encodes for 400 proteins and 200 nm of capsid, comparing to Herpesvirus or HIV that are smaller in terms of genome and viral particle dimensions [28]. As a demonstration of that, the first giant virus discovered was exchanged for a bacterium by GRAM + staining.

Given their unusual properties for a virus, like their morphology, ecology, genome size and gene uniqueness, a new name was proposed for the giant viruses, that is "**giruses**". The semantic and scientific goal of the new name[1] was to emphasize the unique properties of large DNA viruses, which likely represent a unique and shared evolutionary history.

In addition, even if several viral hallmark genes are shared by NCLDVs and other large DNA viruses, such as herpesviruses and baculoviruses, the conservation of the entire set of core genes clearly demarcates the NCLDVs as a distinct class of viruses, as will be fully described in the next pages[43] . According to this findings, more recently, thanks also to the discovery of new species and the demonstrated

evolutionary reconstruction, an official taxonomic rank of the NCLDVs has been proposed, as the order of "Megavirales" referring to the large size of the virions and genomes of these viruses [44].

NCLDVs reserve a well-defined structure inside the cytoplasm where all those mechanisms needed for the viral assembly take place. These structures are defined as "**virus factories**"[45], which are considered a peculiarity of dsDNA viruses that infect eukaryotic organisms. Virus factories are isolated structures in the cell host cytoplasm were all the mechanisms needed for new virion assembly occur. These compartments create a significant evolutional advantage to the virus, because adequate spatial coordination of viral genome replication and assembly, with maximum efficiency in the use of cell resources, and allow viruses to hide from host cell antiviral defenses[46].

The most studied NCLDVs virus factories are the one of Mimivirus, which infects *Amoebae,* and PBCV-1 that infects *Chlorella variabilis*[40,45]. The main goal for the viral infection is the injection of the viral DNA into the host cytoplasm. Mimivirus releases its DNA into the host cytosol after opening the so-called "stargate" and fusing the virion membrane with the one of the phagocytic vesicles. DNA duplication occurs directly in the host cytoplasm and all the processes (replication, transcription and new virion assembly) take place here[40]. On the other hand, it has been recently established that PBCV -1 uses a "bacteriophage-like strategy", where its genome is injected into the cytosol by a spike, while the virion is attached to the host surface. Then, the DNA migrates into the nucleus, where it replicates using the host machinery, and subsequently it is driven to the cytoplasm [41].

As briefly described above, NCLDVs comprise eight families: *Poxoviridae, Asfaviridae, Iridoviridae, Phycodnaviridae, Ascoviridae, Marseilleviridae, Pithoviridae and Mimiviridae*. More focus will be dedicated to *Chloroviruses,* members of the *Phycodnaviridae*, which are object of this work.

1.2.2 Phycodnaviridae: Chloroviruses and Paramecium bursaria Chlorella virus – 1 as a viral model.

The literal translation of *Phycodnaviridae* is "DNA viruses that infect algae"[28]. Viruses infecting eukaryotic algae are huge dsDNA viruses with genomes ranging from 160 to 560 kb with up to 600 protein-encoding genes [47].

Phycodnaviridae are found in aqueous environments throughout the world. They also seem to have a dynamic role in regulating algal communities, such as the termination of massive algal blooms commonly referred to as red and brown tides [47,48]. PBCV-1, which is the most studied *Phycodnaviridae*, is a large, icosahedral, plaque-forming virus that replicates in chlorella-like green algae (for this reason it is classified as a chlorovirus) and its structure, its initial stages of infection and many of its genes resemble bacteriophages[47]. In addition, Chlorovirus hosts are normally symbionts with the protozoan *Paramecium bursaria*, the coelenterate *Hydra viridis* or the heliozoon *Acanthocystis turfacea*[47].

Members of the *Phycodnaviridae* are currently grouped into **six genera** (named after the hosts they infect): *Chlorovirus, Coccolithovirus, Prasinovirus, Prymnesiovirus, Phaeovirus* and *Raphidovirus.* Complete genome sequences have been obtained from representatives of the *Chlorovirus, Coccolithovirus* and *Phaeovirus* genera and evolutionary analysis of their genomes places them as Nucleo Cytoplasmic Large DNA Viruses[28], described before in this chapter.

Despite the wide host range of *Phycodnaviridae,* all members have similar structural morphology, and this is consistent with a common ancestry. Data obtained on PBCV-1 demonstrated that virions are large layered structures of about 100-220 nm in diameter with a dsDNA-protein core surrounded by an icosahedral capsid that covers a single lipid bilayered membrane, which is required for infection.

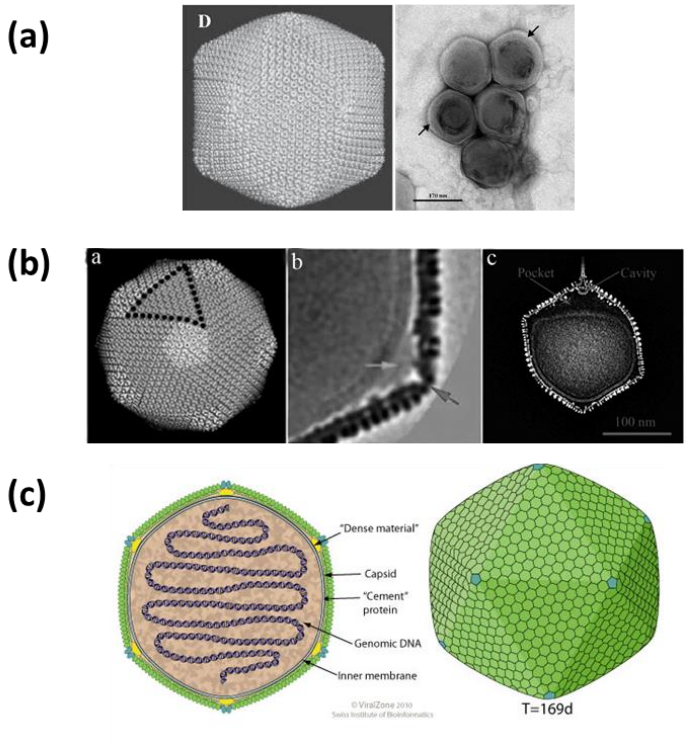


Figure 16. (a) *Chlorovirus (left) and Coccolithovirus (right) as two examples of Phycodnaviridae* (Modified from Van Etten et al., and Wilson et al.). **(b)** *Three-dimensional image reconstruction of chlorella virus PBCV-1*. The virion capsid consists of 11 pentasymmetrons and 20 trisymmetrons (Modified from Stass et al.). **(c)** S*chematic representation of PBCV-1 envelope.* The membrane is surrounded by a capsid with an icosahedral symmetry (T=169), 100-220 nm in diameter (From https://viralzone.expasy.org/145).

As described by Van Etten et al., one of the PBCV-1 capsid vertice shows a long spike-structure that forms a closed cavity inside a large pocket between the capsid and the membrane where the viral DNA is packed. The capsid shell consists of 1680 donut-shaped trimeric capsomers plus 11 pentameric capsomers, one at each icosahedral vertex except for the spike-containing vertex. The trimeric capsomers are arranged into 20 tri- angular facets (trisymmetrons, each containing 66 trimers) and 11 pentagonal facets (pentasymmetrons, each containing 30 trimers and one pentamer at the icosahedral vertices (**Figure 16**). External fibers extend from some of the trisymmetron capsomers (probably one per trisymmetron) and presumably aid in virus attachment to the host [47].

PBCV-1 major capsid glycoprotein is Vp-54, with a predicted mass weight of 48,165 kDa, reaching to 53,79 kDa with posttranslational modifications. It consists of two

eight-stranded, antiparallel β-barrel jelly-roll domains. Vp-54 represents the 40% of the total proteins and identities of the other virus surface proteins are unknown.

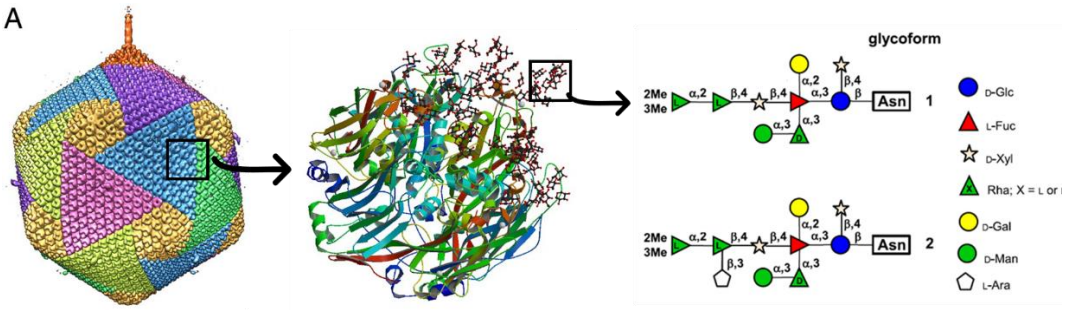


Figure 17. *Representation of PBCV-1 major capsidic glycoprotein and its glycoforms.* Vp-54 is a protein of about 48 kDa and consists of two eight-stranded, antiparallel β-barrel jelly-roll domains. It represents the 40% of the total proteins and identities of the other virus surface proteins are unknown. The two major glycoforms identified so far are N-linked glycans identified only in PBCV-1. (Modified from De Castro et al[49].).

The Vp-54 crystal structure[49] (with PDB code 5TIQ) in **Figure 17** reveals four N-linked glycans attached to the protein, that are recently determined by detailed chemical, spectroscopic, and spectrometric analysis[50,51]. Previous studies incorrectly identified also two O-linked glycans, but without the solved glycan structure, they were just "fixed" in the density X-ray map[49]. Recently the old data have been reviewed, refining these observations and confirming the presence of only the four N-linked glycans[49].

The glycan portion of Vp-54, which is oriented to the outside of the particle, contains seven neutral sugars: glucose, fucose, rhamnose, galactose, mannose, xylose, and arabinose[28]. However, the four glycosylated Asn residues are not located in the typical prokaryotic and eukaryotic N-linked glycans consensus sequences Asn-X- (Thr/Ser)[28,50]. In addition, the structure of the Vp54 glycoforms are unique and do not resemble any type of structure reported so far for any organism in the three domains of life and in NCLDVs too. In fact, the N-glycans of all chlorella viruses have a new structural core that do not resemble any other reported for bacteria, archea or eukarya, thus it can be considered a signature for this class of organisms.

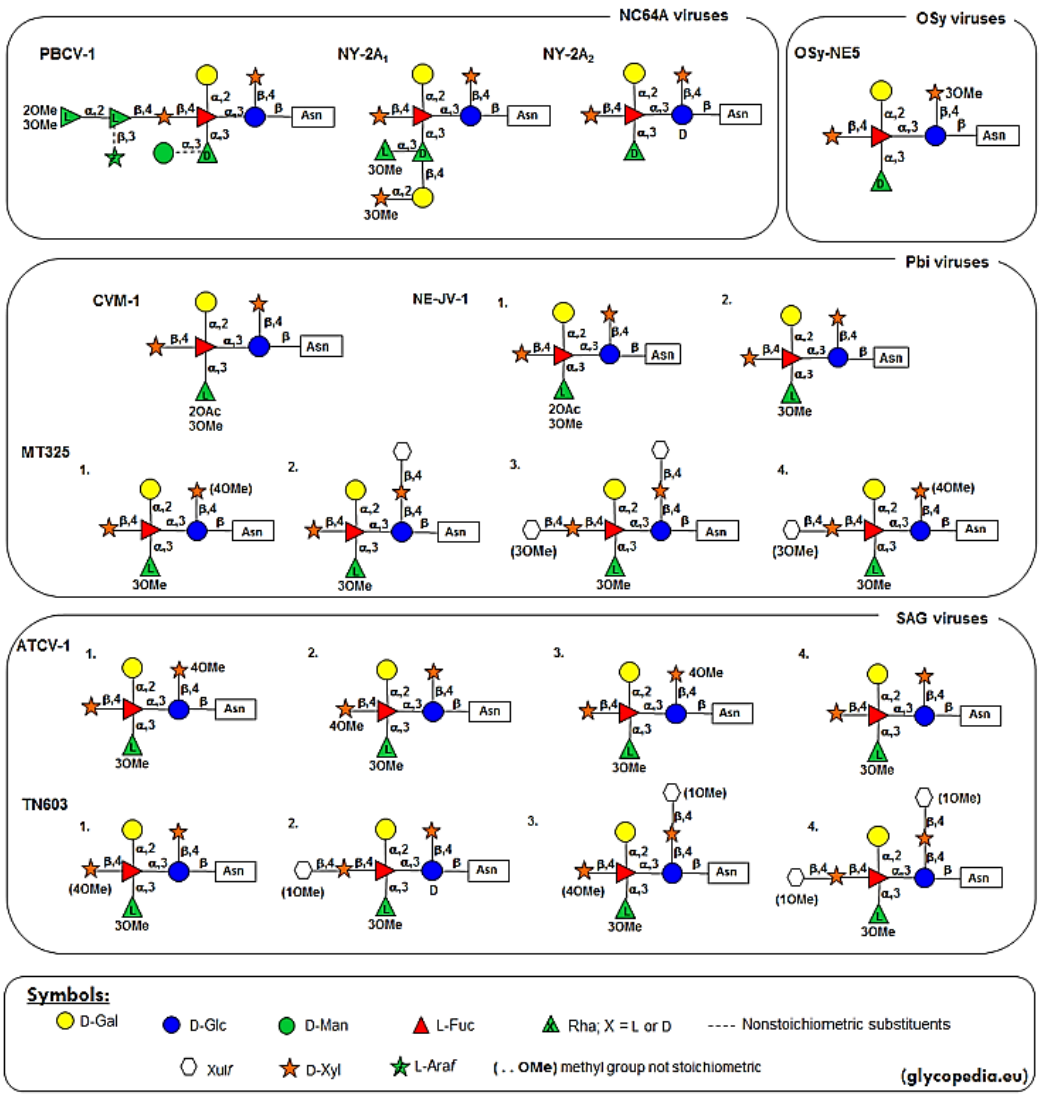


Figure 18. *Different glycoforms among Chlorovirus.* Each glycoform is classified according to the Chlorella host infected (NC64A, OSy, Pbi, SAG). (From Speciale et al[51].)

In particular, the oligosaccharide is linked to the Asn residue by a β-glucose; this type of linkage is very rare, and it is found so far only in some Archea and Bacteria. The oligosaccharides are highly branched, with a fucose substituted at all available positions; each glycan contains two rhamnose residues with opposite configurations (abbreviated as L-Rha or D-Rha) plus a further terminal L-Rha capped with two O-methyl groups (diO-Me-L-Rha). Two monosaccharides, arabinose and mannose, occur as nonstoichiometric substituents, resulting in four glycoforms. Glycoform 1 and glycoform 2, depicted in **Figure 17** are the most abundant[3,49].

Glycoform 1 is a nonasaccharide. It possesses all the monosaccharides except Ara, and it is the predominant form at Asn-302, Asn-399, and Asn-406. Glycoform 2 is a decasaccharide, it includes Ara, and it is the predominant form linked at Asn-280[49].

As **Figure 17** shows, the structure of these N-glycans consists of two regions: the core region is located near the protein backbone and is highly conserved among the chloroviruses[3,49]. It consists of the N-linked glucose, two xylose units [one located close (proximal unit or Xylprox) and one far (distal unit or Xyldist) from the protein backbone], the hyperbranched fucose and a galactose. The second region extends the conserved core with other monosaccharides, which are specific for each chlorovirus[3,49]. All these N-glycans are unique and do not resemble any known eukaryotic or prokaryotic glycan, prompting interest in viral glyco-releated genes[3,49].

In the recent years, this finding together with other observations led to the conclusion that PBCV-1 as other members of NCLDVs [39] encodes at least part, if not all, of the machinery required to glycosylate its major capsid protein independently from the host endoplasmic reticulum-Golgi system [3,28,50]. This is a peculiarity of the NCLDVs that is giving a growing interest to the study of those viruses and that define them as partially independent eukaryotic parasites. In fact, as a support of this thesis, PBCV-1 it is demonstrated to possess own glycosyltransferases that synthetize the glycoforms[52].

As described in a detailed review from Wilson et al, the PBCV-1 virion contains more than 110 different virus-encoded proteins and about 700 ORFs[28]. But, [28]in order to maximise the space inside the virion, viruses typically have compact genomes that help replication efficiency. *Phycodnaviruses* fit this pattern with approximately one gene per 900 to 1000 bp of genomic sequence[28]. The 366 PBCV-1 protein-encoding genes are evenly distributed on both strands and, with one exception (a 1788-nucleotide sequence near the middle of the genome), intergenic space is minimal. In fact, 275 ORFs are separated by fewer than 100 nucleotides. The 1788-nucleotide DNA region, which contains many stop codons in all reading frames, encodes 11 tRNA genes[28].

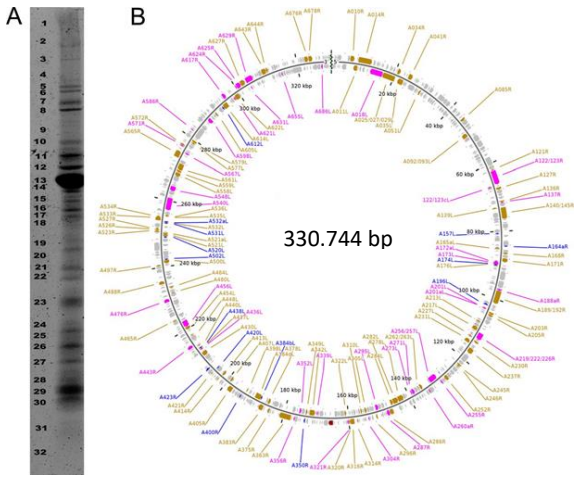


Figure 19. *SDS-PAGE separation and PBCV-1 genome mapping.* (Modified from Van Etten et al[47].).

Some of the *Phycodnaviruses* genomes have methylated bases. For example, genomes from 37 chlorella viruses contain 5-methylcytosine (5mC) and N6-methyladenine (6 mA) that occur in specific DNA sequences. However, it was a surprise to discover that approximately 25% of the virus-encoded DNA methyltransferases have a companion DNA site-specific (restriction) endonuclease. Thus, the virus-infected chlorellae are the first nonprokaryotic source of DNA site-specific endonucleases[28].

Sequencing of *Phycodnavirus* genomes showed genes not previously found in viruses that may provide clues as to their niche adaptation. As an example, it is possible to recognize a sphingolipid biosynthesis pathway in coccolithovirus Ehv-89, the previously mentioned sugar metabolism in PBCV-1, the enzymes involved in the degradation of host cell wall (chitinases, chitosanase, β-l-3 glucanase and enzyme that cleaves polymers of either β-or α-l,4-linked glucuronic acids) and enzymes involved in polyamine biosinthesys found again in PBCV - 1[28].

As previously described, PBCV-1 replication strategy involves host cytoplasm and nucleus, in a partial different way used by *Mimivirus*. PBCV-1 infection starts with the attachment of the virus to the algae cell wall, in an irreversible way (**Figure 20**, panel (a) and (b)). This host-virus recognition is immediately followed by the DNA injection into the host cytoplasm in a bacteriophage-like way and translocation to

the nucleus[41]. The transcription of early genes starts 7 min PI (post infection), and the DNA synthesis 60 min PI in the nucleus, where the first structural evidence of infection is observed: deformed morphologies of host nuclei were detected, as those of infected cells lose their spherical shape and assume elongated or crescent-like morphologies, revealing enhanced heterochromicity (areas near the nuclear membrane that are heavily stained for DNA)[6] (**Figure 20** panel (c) and (d)). Such modified nuclear structures reflect extensive degradation of host DNA[45].

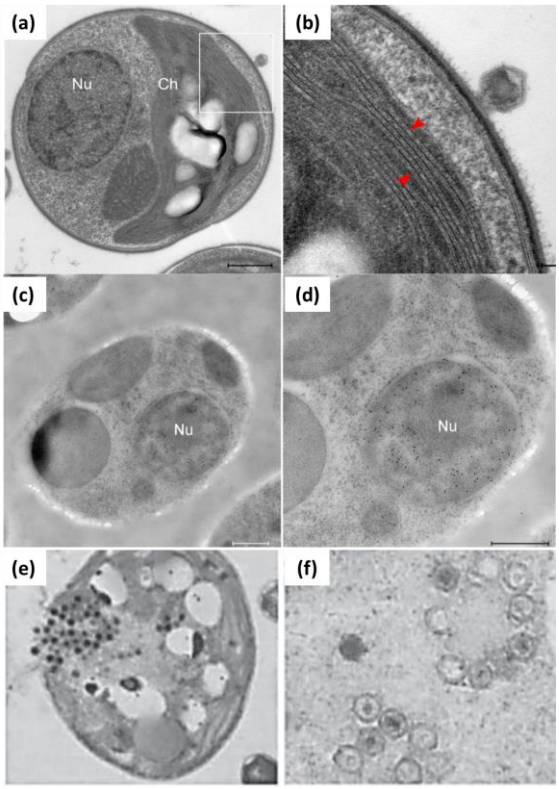


Figure 20. *PBCV-1 major infection stages*. first the virus binds the host cell wall where it injects the DNA. Then, after the DNA translocation to the nucleus, it is driven to the cytoplasm where there is the assembly of the viral factories. After new viral particle assembly, the virus ejects from the cell in a cell litic way (Modified from Milrot et al[41]., and Kang et al[53].).

Differently from the *Acanthamoeba*- infecting *Mimivirus*, Chlorella membranes play a central role in the assembly of the structure of PBCV – 1 cytoplasmic virus factories, detected at 2h PI (**Figure 20**, panel (e) and (f)). Sequential tomography slices in **Figure 20** show early viral factories as rosette-like crescent structures that consist of two distinct layers characterized by different densities: an external

angular capsid shell and an internal membrane bilayer, where host ribosomes and other organelles are excluded, probably for the massive accumulation of new progeny virions[45]. Factory generation is sided by massive accumulation of the host membrane cisternae that partially surround the viral factories and, in some cases, deeply penetrate the factory cores. These cisternae appear to bud out from rough endoplasmic reticulum (ER) membranes and they are derived mainly from outer membranes of host nuclei[41]. In contrast to *Mimivirus* and *Vaccinia* virus viral factories, cores of PBCV-1 factories consist of membrane structures surrounded by viral genomes. Moreover, in the viral factories, it is possible to see the different virion assembly stages starting from crescent-shaped structures, partially assembled particles lacking DNA and mature virions. Mature virions are forced away from the viral factory core presumably by the progressive and continuous generation of new progeny viral particles[45] (**Figure 20** panel (e)**).** At 3h PI the host cytoplasm appears to be full of viral DNA that is located at viral factory periphery and, at the same time, host nuclei is completely empty of host DNA, in agreement with the demonstration that host DNA is degraded after virus infection by viral endonucleases, and the synthesis of new viral particles is driven by viral proteins[45]. Other members of *Phycodnaviridae* are not well studied as PBCV-1, but of particular interest is the *Coccolithovirus* with its mentioned spingolipid byosinthesis pathway. Coccolithovirus infects a marine microalgae with global distribution, identified as *Emiliana huxley*[28]. The prototype specie of the genus *Coccolithovirus* is EhV-86 and, as all the *Phycodnaviridae* and NCLDVs, it possesses a large genome[54].

The *Coccolithovirus* sphingolipid biosynthesis pathway is important for its infection strategy that involves unique mechanisms for replication, survival, defence, evolution, dissemination, and communication[55]. In particular, it is demonstrated to possess genes involved in the synthesis of ceramide, a sphingolipid that induces apoptosis[55].

1.3 Glycosylation in NCLDVs

As largely described in the previous chapter, NCLDVs display some eukaryotic characteristics that are uncommon in viruses. In addition to the large dimensions in terms of DNA and viral particle, NCLDVs possess proteins that allows them to be partially independent from the cell host and permit to drive all the host resources to feed their pathways[25]. The viral particle assembly take place in defined zones in the host cytoplasm known as “virus factories”, described above, where possibly the viral protein glycosylation take place[39].

In fact, together with genes involved in DNA replication, protein synthesis and different type of posttranslational modifications, genome sequencing of NCLDVs also revealed the presence of several genes involved in glycosylation, including glycosyltransferases and other carbohydrate-modifying enzymes. In addition, as described by Piacente et al., genomes of large DNA viruses often present enzymes responsible for the production of the nucleotide-sugars that are the substrates for glycan formation [39].

1.3.1 Chlorovirus glycosylation and nucleotide sugar biosynthetic pathways

Several nucleotide-sugars biosythetic pathways have been identified in NCLDVs. The GDP-L-fucose pathway was first described in PBCV-1 by Tonetti et al in 2003[52]. Since then, enzymes involved in the hexosamine pathway, D- and L-rhamnose and D-viosamine have been found[56]. Moreover, genome inspection of most reported members of the Phycodnaviridae reveals the possible presence of many other putative enzymes involved in nucleotide-sugar production.

PBCV-1 encodes two enzymes involved in the nucleotide-sugar biosynthetic pathways: a GDP-D-mannose 4,6-dehydratase (**GMD**, a118r gene), and a NADPH-dependent bifunctional GDP-4-keto-6-deoxy-mannose epimerase/reductase (**GMER**, a295l gene). GMD produces the unstable intermediate GDP-4-keto-6-deoxy-mannose from the GDP-D-mannose, and GMER which catalyses the 3,5-epimerization followed by a NADPH-dependent reduction of C-4[39] (**Figure 21**).

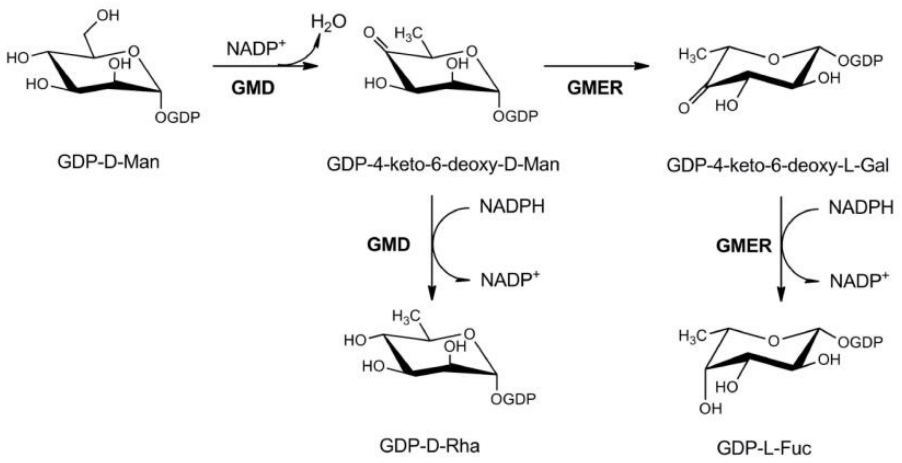


Figure 21. *Metabolism of GDP-D-rhamnose and GDP-L-fucose in PBCV-1*. Both GMD and GMER have two activities: PBCV-1 GMD with both dehydratase and NADPH-dependent reductase activities, leading to GDP-D-rhamnose formation. GMER is a GDP-4-keto-6-deoxy-D-mannose 3,5 epimerase/4-reductase producing GDP-L-fucose[39].

BLAST analysis indicates that both GMD and GMER are conserved in most *Chloroviruses* sequenced so far, suggesting an essential role of this pathway in viral replication. Actually, GDP-L-fucose is produced by chlorella hosts, however, the cytosolic amounts of this nucleotide are usually low and GMD, the limiting step of the pathway, is subjected to strong feed-back regulation by its products. Indeed, L-fucose is an important component of the PBCV-1 glycan core structure and, as consequence, the pathway is probably used to circumvent a limited supply of GDP-L-fucose, which might limit the oligosaccharide synthesis[39].

In PBCV-1 and in some other closely related Chloroviruses, GMD is also a bifucntional enzyme, displaying also a NADPH-dependent reductase activity on C-4 of the intermediate compound obtained through its dehydratase reaction. This leads to the formation of GDP-D-rhamnose. This bifunctional activity of PBCV-1 GMD is not observed for the enzyme encoded by other Chloroviruses, such as ATCV-1 [30] . This is consistent with the finding that D-rhamnose is a component of the variable region of PBCV-1 glycan, but it is not found in ATCV-1 and most other members of this genera [39]

UGD, the UDP-D-glucose dehydratase that is the first enzyme of the L-rhamnose pathway, is found only in ATCV-1 and other few isolates, all infecting *Chlorella heliozoae*, the endosymbiont of *Acanthocystis turfacea.* The second enzyme of the pathway, a bifunctional epimerase/reductase is absent in ATCV-1. In fact, UDP-L-rhamnose is commonly produced by plants and L-rhamnose synthesis has been demonstrated to occur in Chlorella algae. Since UGD represents the limiting step for UDP-L-rhamnose production, it is possible to hypothesize that the enzyme was acquired by the virus to prevent feed-back inhibition and increase the nucleotide-sugar supply. In addition, it has to be noted that genes encoding UGD-like proteins are not present in sequenced chloroviruses infecting other hosts, suggesting that this enzymatic activity is not essential for all chlorella viruses[39].

PBCV-1 and several other chloroviruses induce the formation of an extracellular polysaccharide, either hyaluronan or chitin, a short time after infection. Two enzymes involved to overcome a limited supply of the two precursors for the hyluronan biosynthesis are found in PBCV-1 genome: a100r gene product is a functional glutamine-fructose-6P aminotransferase (**GFAT**), which catalyzes the

first step for the de novo UDP-D-N-acetylglucosamine (UDP-D-GlcNAc) pathway, and a609l gene product, that is a UDP-D-glucose dehydrogenase (**UGDH**), which promotes the NAD^+-dependent oxidation of glucose C-6, leading to the UDP-D-glucuronate formation[39].

PBCV-1 genome has been demonstrated to encode at least 6 glycosyltransferases genes: A064R, A111/114R, A075L, A546L, A219/222/226R, A473L (**Figure 22**) .[2] These glycosyltransferases are predicted to be soluble proteins and located in the host cytoplasm[51]. Interestingly, the six encoded glycosyltransferases of PBCV-1 do not justify the complexity of its glycan structures; an explanation is that some of the glycosyltransferases display more than one GT domain. Another possibility is that viruses encode for proteins that cannot be predicted as glycosyltransferases just by sequence comparison, so they could be overlooked during the data base searches. In addition, some host glycosyltransferases could be recruited for the glycan synthesis[51].

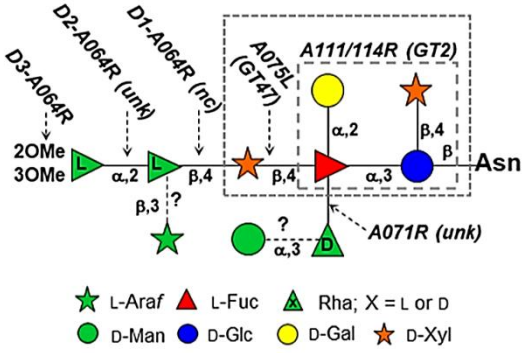


Figure 22. *PBCV-1 Vp-54 major representative glycoform*. On the glycoforms are signed the putative glycosyltransferases suggested to be responsible of the synthesys: A064R, A075L, A071R and A111/114R[51].

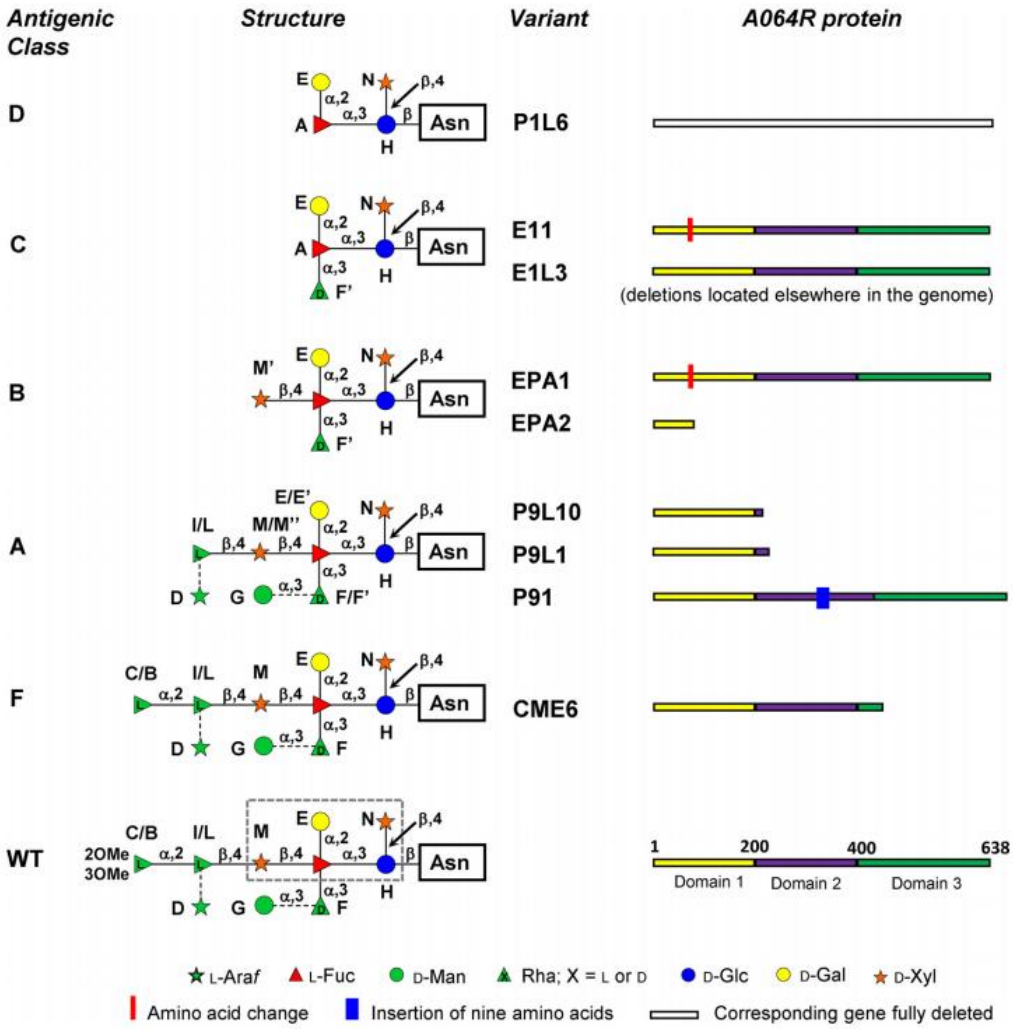


Figure 23. *Glycan structures of WT and selected antigenic mutants of PBCV-1*. Antigenic classes are labeled with a capital letter (A–D and F), and the structure of their glycans depends on the kind of mutation found in the genome, which in all cases, except for Classes C and D, occurs in the gene a064r. EPA1 (antigenic Class B) and E11 (antigenic Class C) have identical mutations in domain 1 of a064r, but E11 has an additional mutation in gene a075l. P91 has an additional glycoform with -L-Rha methylated at O-2 (From Speciale et al[51].).

To date, the only gene products for which the enzymatic activities have been confirmed are PBCV-1 hyaluronan synthase (HAS, a98l gene product) and CVK2 chitin synthase (CHS). In addition, two chlorovirus putative glycosyltransferases have been crystallized, a064r N-terminal domain gene product from PBCV-1 and b736l gene product from the related NY2A strain. The A064R N-terminal domain shows a GT-A fold [57] (PDB code: 2P73) similar to retaining glycosyltransferases, while the C-terminal domain resembles an O-methyltransferase. Indeed, the a64r

gene is restricted to few viral species and it is even absent in viruses closely related to PBCV-1.

The second chlorovirus glycosyltransferase for which the structure has been determined is b736l gene product from PBCV-NY2A virus infecting *Chlorella variabilis* NC64, that is well conserved in viruses infecting *Chlorella variabilis*[39].

The features of the putative PBCV-1 glycosyltransferases have been determined also thanks to the study of Vp-54 glycoforms and the identification of mutant variants, represented in **Figure 23**[51]. As reported by Speciale et al[51], PBCV-1 spontaneous mutants are divided in to six antigenic classes denoted with a letter, based on their differential reaction to five different polyclonal antibodies. An exception is Class E that cross-react with Class A and B polyclonal antibodies, suggesting that the phenotypes of class E variants shares some of the structural features of the other two variants[51]. It is important to note that five of the six classes have mutation in the A064R gene. Indeed, the mutants lacks some sugars of the N-glycan moieties[51].

Secondly, another approach for the identification of PBCV-1 glycosyltransferases is based on genome comparison of the genomes of all *Chlorovirus* identified so far[2]. As already mentioned, all *Chlorovirus* glycans share a common core structure. With this approach, some genes encoding enzymes possibly involved in the glycan core formation have been identified, including a111/114r and a075l[2].

A111/114R is a protein of 860 amino acids organised into three domains, where the second one is annotated as a glycosyltransferase. A111/114R has a possible role in the core glycan assembly and attachment. Indeed, this gene is not affected by the large genomic deletion of the serological mutants reported in **Figure 23**. Since it is the only annotated orthologous glycosyltransferase gene found outside the deletion regions in mutants, it is present in all Chloroviruses and it was not possible to identify viable mutants for the gene, this led to the conclusion that is responsible for the formation of the core N-glycan and that this portion is the minimal required for virus viability, even if its role is still unclear[51].

According to the mutant analysis, A071R is supposed to attach the α,3 D-rhamnose to the α,3 fucose[51]. In addition, it is classified as a glycosyltransferase even if does not displays any resemblance in the database, except of domain 2 of A064R that

is also not been annotated but there is strong evidence that it is a glycosyltransferase (see results section)[51].

Moreover, by genetic analysis of the mutants, A075L is supposed to attach the distal xylose. The evidence came from the N-glycan variants from Class C, that lacks this xylose and display a mutation in the a075l gene[51]. With the same kind of observation, A064R domain 1 and 2 are classified as rhamnosyltransferases[51].

A064R variants displaying mutations in the C-terminal domain lack the methyl groups on the last L-rhamnose, suggesting the enzymatic activity of the protein, and are classified in the Class F genetic variants. Even if PBCV-1 encodes many DNA methyltransferase genes, the only mutated methyltransferase gene in the antigenic variant is in the C-terminal domain a064r. The best hit in a BLAST search of the WT amino acid sequence of domain 3 to the nucleotide database at NCBI was to 2'-O-rhamnosylmethyltransferases. Indeed, when the rhamnose methyltransferase from *Mycobacterium smegmatis* was BLASTED against the PBCV-1 protein database, the only hit was to domain 3 of A064R (33% identity and 49% positive). Consequently, when a064r is mutated in the third domain, the corresponding N-glycan is not methylated at either the O-2 or O-3 position, suggesting that this methyltransferase can perform a double methylation of the monosaccharide. However, we cannot exclude the possibility that domain 3 attaches only one methyl group and that a second methyltransferase, yet to be identified, is required to complete the full decoration of this rhamnose[51].

1. Experimental Procedures

1.1 Gene cloning in pGEX-6P1 vector

All the sequences are obtained from NCBI: the sequence identifier for A064R is AAC96432.1 with Gene ID: 918349 and for A075L AAC96443.1 with Gene ID: 917877. Sequences were analysed using the BLAST tool.

The full length A064R gene, cloned in pDEST42-V5-His vector, was kindly provided by Professor Van Etten from the University of Nebraska. Primers were designed to clone each domain separately in pGEX-6P1, by restriction cloning with BamHI and XhoI enzymes (NEB). A scheme of the cloned sequences is shown in the Results section. The Q5 High fidelity DNA polymerase (NEB) was used to amplify the fragments of interest, following the supplied protocol: initial denaturation 98°C for 30 seconds, [98°C 10 for seconds, annealing temperature (see **Table 3**) for 30 seconds, 72°C for30s/kb] repeated for 35 cycles, final extension at 72°C for 2 minutes. The annealing temperature for each fragment is displayed in **Table 2.**

Primers table

GENE NAME		PRIMER SEQUENCE	PCR PRODUCT LEGHT (bp)	RESTRICTION ENZYMES	Tm °C
A064R D1 + D2 short	F	AATT **ggatcc** ATGACCACACCTTGTATTAC	1230	BAM HI	39
	R	AATT **ctcgag** TTAATTTTTGTACACAGGGT		XHO I	37
A064R D1 + D2 long	F	AATT **ggatcc** ATGACCACACCTTGTATTAC	1329	BAM HI	39
	R	AATT **ctcgag** TTAGGTTTCCTCCGTCGAGG		XHO I	45
A064R D1	F	AATT **ggatcc** ATGACCACACCTTGTATTAC	648	BAM HI	39
	R	AATT **ctcgag** TTAAGTTGCTACCATCTCCA		XHO I	40
A064R D2 short	F	AATT **ggatcc** ATGTGCGGTACTTCTCGTGC	651	BAM HI	47
	R	AATT **ctcgag** TTAATTTTTGTACACAGGGT		XHO I	37
A064R D2 long	F	AATT **ggatcc** ATGTGCGGTACTTCTCGTGC	750	BAM HI	47
	R	AATT **ctcgag** TTAGGTTTCCTCCGTCGAGG		XHO I	45
A064R D2 long 2	F	AATT **ggatcc** ATGGTAGCAACTGGTGAAAT	714	BAM HI	41
	R	AATT **ctcgag** TTAGGTTTCCTCCGTCGAGG		XHO I	53

Table 2. Primers used to clone A064R domains

NAME	Ta °C
A064R D1 + D2 short	52
A064R D1 + D2 long	54
A064R D1	54
A064R D2 short	52
A064R D2 long	60
A064R D2 long 2	56

Table 3. *PCR condition for A064R gene domains.*

A075L gene was amplified using PBCV – 1 DNA, kindly provided by professor Van Etten, University of Nebraska. PCR amplification was performed as indicated above. Primer sequence for PCR amplification is reported in **Table 4**.

GENE NAME	PRIMERS SEQUENCE		PCR PRODUCT LENGTH (bp)	RESTRICTION ENZYMES	Tm °C
A075L	F	AATT **gatcc**ATGAAGCTCGCCGAACTTAC	810	BAM HI	44
	R	AATT**ctcgag**TTACTGACTATATTCGAGAA		XHO I	34

Table 4. *A075L primers table.*

Cloning procedure into pGEX-6p1 vector expression system

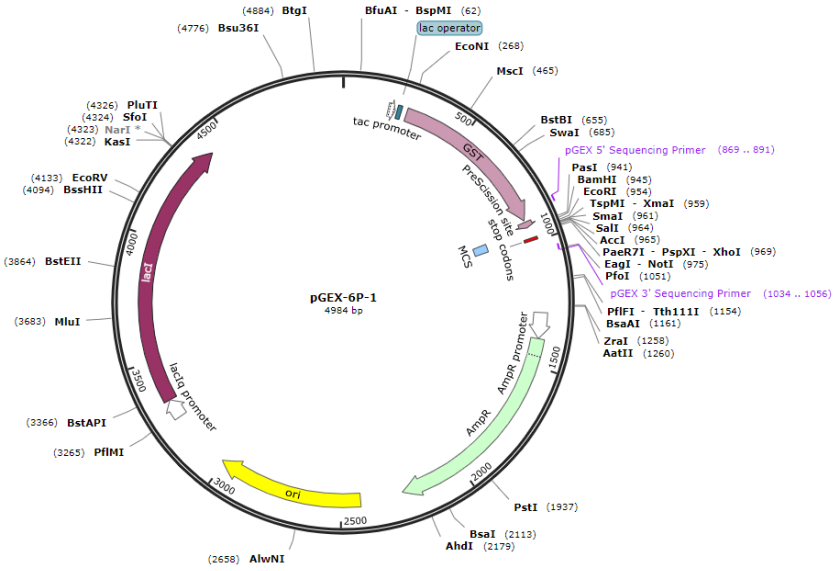


Figure 24. *pGEX-6p1 vector map from snapgene.com.*

pGEX-6p1 is a prokaryotic expression vector: it displays a 4984 bp sequence that contains ampicillin resistance and the N-terminal GST (Gluthatione S- transferase) tag. GST is a small protein of 25 kDa that, thanks to its solubility, helps the recombinant protein purification. In addition, it is also possible to monitor the protein purification procedure with the CDNB assay that, in the presence of glutathione (GSH) and 1-chloro-2,4-dinitrobenzene (CDNB), allows to follow the kinetics of increase of absorbance at 340 nm. In addition, the protein of interest is recovered, after the affinity binding to the GSH-resin, by a specific proteolytic site between GST and the protein of interest, using the Prescission Protease (GE Healthcare). Prescission Protease is a genetically engineered fusion protein of human rhinovirus 3C protease and GST that specifically cleaves between the Gln and Gly residues of the recognition sequence LeuGluValLeuPheGln/GlyPro.

After PCR amplification by the high-fidelity DNA polymerase, the correct length of the products was analysed by 1.5% agarose electrophoresis. For A064R, for which the pDEST-V5-His vector was used as template, a further incubation with 1 U/50µl of DpnI enzyme (Roche) was performed for 1 hat 37°C after the end of PCR cycles,

in order to ensure complete removal of the template vector. PCR products were then purified using the “PCR clean-up kit” (Sigma). After purification, the amplified DNA was subjected to restriction digestion for 1 h at 37°C, using BamHI and XhoI enzymes (NEB). The products were purified again after the end of incubation, by the “PCR clean-up kit”.

The pGEX-6p1 vector was digested using the same protocol. Complete digestion of the vector was verified by agarose electrophoresis. A following dephosphorylation step, using Antarctic phosphatase (NEB), was added after digestion. Restriction enzymes and phosphatase were inactivated as indicated by manufacturer’s instructions.

Ligation was performed using the “Quick transformation kit” (Roche), following the provided instructions. The ligation product was also analysed by 1% agarose electrophoresis.

TOP10 *E. coli* cells were transformed with 2 µl of the ligation product, directly added to 100 µl of the competent cells: cells were then incubated for 30 min on ice. Then, to facilitate vector entry, heat shock at 42°C was applied for 45’’, followed by 2 min on ice. SOC medium (1 ml) was added and cells were incubated on a rotary shacked for 1h at 37°C. The cell suspension was then centrifuged at 2500 rpm for 10’ and resuspended in 300 µl of SOC medium, which was then spread on LB agar plates and incubated at 37°C overnight.

The day after, some colonies were chosen to be verified by PCR followed by 1,5% agarose electrophoresis of the amplification product. One colony that resulted positive for the right insert was then grown in 5 ml medium overnight. The vector was purified with the Miniprep kit from Sigma and sequenced. The verified vector was transformed into BL21 competent cells, using the procedure describe above, with a heat shock at 42°C for 30’’, for expression of the recombinant protein. Full length A064R cloned in pDEST64-V5-His was transformed in BL21-DE3 cells, using the protocol described above.

1.2 A064R

1.2.1 Expression and purification of the recombinant proteins: A064R full length and A064R domains

Bacterial growth and lysate preparation

A 10 ml starter culture was grown overnight at 37°C. The day after, it was diluted 1:100 in fresh medium; cells were grown for approximately 8 hours at 20°C until OD_{600} was 0.4 – 0.6. Protein expression was induced with 0.1 mM IPTG (Sigma) overnight, in vigorous shaking at 18°C, with humid air bubbling.

The bacteria were recovered by centrifugation at 7000 rpm for 15 min, then the pellet was re-suspended in lysis buffer. For the full length A064R, which is produced with a C-terminal 6XHis tag, 50 mM sodium phosphate buffer containing 500 mM NaCl, pH 8.0, 5 mM imidazole, was used for cell resuspension. Conversely, E. coli cells expressing the GST-fusion domains were re-suspended in PBS (50µl /L of the initial culture volume). Cells were disrupted by sonication (10 cycles of 10s on/10s off repeated 3 times) and, for the proteins produced with the GST tag, 1% Triton X-100 was also added and incubated for 30 min at 4°C with gentle agitation, to promote protein release from cell debris. Then, the insoluble fractions were removed by centrifugation at 12 000 x g for 25 min.

Purification of recombinant proteins: 6xHis tag expression system (A064R full length)

The supernatant obtained after lysis and the last centrifugation step was incubated in batch with 1 ml with a nichel affinity resin (Probond) for 350 ml of the initial bacterial culture: beads were pre-equilibrated with binding buffer (50 mM Na phosphate buffer, 500 mM NaCl, 5 mM imidazole, pH 8.0) for 1 h at 4°C with gentle agitation. At the end of this incubation time, the beads were packed in a 15 ml column and the flow through, which corresponds to the unbound proteins, was collected for further analysis. The column was washed with 10 volumes of wash buffer (50 mM Na phosphate, 500 mM NaCl, pH 8.0, 20 mM imidazole); then the protein was eluted with wash buffer supplemented with 250

mM imidazole. The eluted fractions containing the recombinant protein were recovered and concentrated using the Millipore Centrifugal units, cut off 10 KDa, by centrifugation at 2800 rpm, until a final volume of 500 µl was reached.

Purification of recombinant proteins: GST tag expression system, A064R gene domains

The supernatant was initially tested for GST activity, following manufacturer's instructions, and it was incubated with 1 ml of GSH Sepharose (GE Healthcare) for 350 ml of initial bacterial culture, to affinity capture the protein of interest. The binding to the beads was carried for 1 h at 4°C in gentle agitation. The beads were then packed into a 15 ml column and the flow through devoid of the GST-fusion protein was recovered for further analysis. The GST assay was repeated on the eluate, in order to calculate the extent of binding of GST-fusion protein the resin. The beads were washed with 10 volumes of PBS, then with 5 volumes of cleavage buffer (150 mM NaCl, 50 mM Tris/HCl pH 7.5) and finally they were recovered from the column with 15 ml of cleavage buffer; the protein of interest was then cleaved from the GST-tag, by addition of 1 mM DTT and of Prescission Protease (5µl/ml of GSH-Sepharose used; GE Healthcare) in batch, at 4°C overnight in gentle agitation.

The day after, the resin was again packed in a 15 ml column and the flow throw with the recombinant protein was recovered and concentrated using the Millipore Centrifugal units, by centrifugation at 2800 rpm, until a final volume of 500 µl was reached.

For some experiments, the GST-fusion proteins were not cleaved and released from the resin, but after the last wash the immobilized proteins were directly tested for the enzymatic activity.

1.2.2 Enzymatic Characterization

D1 and D2 are putative UDP-rhamnosyl transferases. These domains, as it will be detailed in the Results section, were cloned separately or together; for the second domain, different length of the sequence was cloned, in order to define the minimum sequence needed for the enzymatic activity.

To test the enzymatic activities, in collaboration with Professor Cristina De Castro from the University Federico II in Naples, a synthetic acceptor named "Todd" was produced by Prof. Todd Lowarty at the University of Alberta. This compound displays the distal xylose of the PBCV-1 glycoform, as shown in **Figure 29** in the Results section. Both activities were tested in the presence or absence of cations or of chelating agent EDTA. The reaction mixtures (1.5 mM "Todd", 1.5 mM UDP-L-Rha, 2 mM Mg^{2+} or Mn^{2+}) were carried out in PBS overnight (O.N.) at 30° C with the protein still bound to the resin beads; samples were collected at T_0, T_{4h} and $T_{O.N}$. The conversion of the acceptor substrate to the product was verified by HPLC using a C18 column with 70% MeOH as eluent (flow rate at 0.8 mL/min; 10 µL: volume of each injection). The eluate was monitored by a refractive index detector.

The activity of **D3** domain is supposed to be a methyltransferase able to attach the distal methyl groups on the last L-rhamnose residue. This domain has not been produced alone as yet; however, preliminary tests were performed incubating the full-length protein with "Todd" as acceptor and a fivefold excess of S-Adenosyl-methionine (Sigma), with the protein bound to the resin beads.

To conform the identity and linkage configuration of the products of D1, D2 and D3, NMR analysis was performed on the overnight reaction, after a Sep-Pak column purification to remove cations. This step was essential to remove especially manganese which is paramagnetic and prevents products analysis via NMR. The Sep-Pak cartridge was first activated, following manufacturer's instructions; the scheme of elution was: 20 mL of water; 20 mL of acetonitrile/water in 1:4 ratio, 8 mL of acetonitrile, 20 mL of ethanol.

All proton NMR were recorded in D_2O at 310 K on a Bruker DRX-600 MHz instrument equipped with a cryo-probe. Standard Bruker software, Topspin 3.1, was used for process and analysis of all spectra.

1.3 A075L

1.3.1 Expression and purification of the recombinant protein

For A075L purification, slight modifications were added to the GST standard protocol in order to increase protein recovery and purity, that is a mandatory condition for X-ray crystallography experiments. In addition, A075L was also produced in a selenomethionine enriched media (SeMet A075L) in order to obtain the substitution of Met residues in the sequence. This step is necessary to solve the 3D structure after the x-ray diffraction. Expression, purification and further analyses of the protein were performed at the Cristallography platform under the supervision of Dr. Adriana Rojas, at CiC Biogune in Bilbao, Spain.

Bacterial growth

An overnight culture was diluted 1:100 in six flasks with 2 litres of Luria Bertani (LB) media each. As the pGEX-6p1 uses Ampicillin as selection marker, 100 µg/ml of this antibiotic were added to the media. The growth was carried at 37°C until an OD_{600} 0.6 was reached, then protein expression was induced with the addition of 0.5 mM IPTG overnight at 18°C.

For the production of the SeMet protein, cells were initially grown in the LB media, (4 liters total), then bacteria were recovered by centrifugation when the OD_{600} reached 0.6, and washed twice with ice cold PBS and placed in a minimum media (SeMet Medium Base (Sigma) supplemented with nutrients (Molecular Dimension) and without methionine, at 37°C for 1h. Then the protein expression was induced with 0.5mM of IPTG overnight at 18°C and Se-Met at final concentration of 40 µg/ml was added. After protein purification describes below, SeMet A075L is verified by MALDI-TOF in order to confirm the incorporation of the SeMet in the protein. For the MALDI-TOF a C4 microcolumn desalting method was used prior to analysis, with 70% acetonitrile and 30% of H_2O as eluent.

Bacterial cells were recovered by centrifugation at 7000 x g for 25 minutes, then the pellets from 4 litres culture were re-suspended in 150 ml of PBS. Cells were

lysed by sonication (10s of burst and 59 sec off. 8 min with 60% of amplitude), then cell homogenate was clarified by centrifugation at 20 000 x g for 40 min.

Purification of the recombinant protein

About 1ml of Glutathione Sepharose 4B resin (GE Healthcare) for each liter of culture medium used for bacterial growth were pre-equilibrated with PBS; the clarified supernatant was incubated with the beads for 1.5 hour at 4°C in gentle agitation. The beads were then washed with 500 ml of PBS and, subsequently, the fusion protein GST A075L was eluted in batch with 4 column volumes of 20 mM GSH in PBS, at room temperature, for 20 minutes. Proteolytic cleavage of the GST tag and removal of GSH used for elution were performed together: briefly, 50 U/ µl of Prescission protease were added to the eluted fusion protein with 1 mM DTT and the resulting solution was dialyzed overnight at 4°C against 1 L of cleavage buffer (50 mM Tris-HCl, 150 mM NaCl, 1mM DTT).

The day after, the solution containing the recombinant A075L released from GST was incubated in batch for 2h at 4°C with 7 ml of GSH beads, using gentle agitation, in order to capture the cleaved GST and the Prescission protease. At the end of the incubation time the beads were transferred in a 50 ml glass column (BioRad) and the flow through containing A075L was collected. This procedure was performed twice, to ensure complete removal of the GST- containing contaminants.

To remove other possible contaminating proteins, A075L was diluted 3 times with Q buffer A (50 mM Tris-HCl, pH 7.5, 1 mM DTT) and it was then loaded in a 5ml Q HP (GE Healthcare) anion exchange column using an AKTA FPLC system. Elution was carried on with a gradient from 0% to 50% of Q buffer B (1M NaCl, 50 mM Tris-HCl pH 7.5, 1 mM DTT) in 10 CV (column volumes). Protein elution was monitored at 280 nm. Fractions containing A075L protein were pooled together and concentrated with centrifugal Millipore filter units (cutoff 10 KDa) before proceeding for further analyses.

1.3.2 Enzymatic characterization

Since A075L is a putative UDP-xylosyl transferase, the enzymatic activity was tested in the presence of UDP-D-xylose (Carbosynth) and different acceptors. Initially, L-fucose was used as single monosaccharide and also as octyl-fucose. However, no activity could be observed in these conditions. As consequence, preliminary experiments were done using the Vp54-associated glycan of E11 PBCV-1 variant. This variant (see introduction section) lacks the two distal L-rhamnose residues and the distal xylose. The glycopeptide was obtained after isolation of Vp54[58], by thermolysin digestion and purification to obtain the pure glycopeptide. Due to the very low amounts of the glycan acceptor, the reaction was carried out in the presence of the cations Mg^{2+} and Mn^{2+} and with the protein in the soluble form.

1.3.3 A075L Substrate Binding reactions

The substrate binding reactions for A075L were performed at CiCBiogune in Bilbao(spain) with the collaboration of Professor Jesus Jimènez Barbero, and the crystallization experiments were also performed at the crystallography platform of CiCBiogune, under the supervision of the platform manager Dr. Adriana Rojas.

Isothermal Titration Calorimetry ITC

To demonstrate that A075L is effectively able to bind UDP-Xylose as a substrate, Isothermal Titration Calorimetry (ITC) was used. ITC technique is highly sensible and allows to understand if the active site of an enzyme is able to interact with the putative substrate, measuring small variations of the temperature (endothermic or exothermic) due to the binding of the molecule in the active site.

The instrument consists of two cells. One cells contains the macromolecule, and in the second cell the ligand is injected with a syringe. Both cells are kept at steady temperature and pressure [57].

When the ligand solution is injected into the cell, the ITC instrument detects heat that is released or absorbed as a result of the interaction. This is done by

measuring the changes in the power needed to maintain isothermal conditions between the reference and the sample cell.

The heat change is calculated by integrating the power over the time (seconds) that gives the enthalpy of the reaction. The heat discharged or consumed during the calorimetric reaction corresponds to the fraction of bound ligand. The increased ligand concentration leads to saturation of substrate and finally less heat is discharged or consumed [57].

Injections are performed repeatedly, and they result in peaks that become smaller as the biomolecule becomes saturated. If no interaction exists between the molecules, the peak sizes remain constant and represent only the heat of dilution. Once titration is completed, the individual peaks are integrated by the instrument software and presented in a Wiseman plot. An appropriate binding model is chosen, and the isotherm is fitted to yield the binding enthalpy ΔH, the K_D and the stoichiometry. The main aspect that needs to be considered, is the appropriate concentration of ligand and protein.

All ITC measurements were carried out at 25 °C on a Microcal PEAQ-ITC (Malvern). The ITC data were processed using Origin software (OriginLab Corp., USA). Before the ITC analysis, A075L protein (155 µM) was dialyzed overnight in phosphate buffer with 1 mM $MnCl_2$ or $MgCl_2$; the next day the UDP-xylose was also diluted in the same buffer.

For these experiments the ligand (1mM UDP-xylose) was titrated with the protein used at 100 µM concentration. In total 19 injections with spacing of 180s were performed; the first injection of 0.4 µl was not used in data fitting, the following injections contained 2 µl of 1 mM UDP-xylose.

Nuclear Magnetic Resonance

NMR instrument (Buckner 600) was used in order to confirm the binding between substrate and protein. The reactions were set up using 45 µM of protein and 1mM of UDP-xylose in the presence or absence of cations at 1mM and also with the addition of 0.5 mM EDTA. The measurement was carried at 25°C for 2h and overnight.

Finally, a Saturation Transfer Difference (STD) experiment was performed in order to understand the ligand binding epitope, the part of the ligand in contact with the protein. From NMR data we calculate the % of STD as an estimation of protein-ligand proximity. The reaction was performed with 45µM A075L, 1mM $MgCl_2$ and 1mM UDP-Xylose.

1.3.4 A075L Crystallization procedure

Crystallization is a process by which atoms or molecules are highly organized into a structure known as a crystal. The principle that influences the crystal formation is the searching of parameters that influences such process. Then, this multiple set of factors allows to yield the crystal[59].

A crystal is a solid composed of a regulatory repeated arrangement of atoms defined as unit cell. When the structure of one unit cell is clarified, the entire structure will be understood. Then, the X-ray diffraction briefly described below, gives complete information of the unit cell dimensions[59].

Protein crystals are made of approximately 50% solvent (that can vary from 25% to 90%) and the protein that occupies the remaining volume. The entire crystal is consequently an ordered gel permeated by extensive interstitial spaces through which solvent and other small molecules can diffuse[59]. The pursuit of protein crystallisation usually requires the identification of chemical, biochemical and physical conditions that allows to yield some crystalline material, and the systematic alteration of these conditions that allows to obtain optimal samples for diffraction analysis. First, the crystallisation conditions are set up as a systematic variation of the most important variables such as pH conditions, precipitants etc. then, once some crystals (or microcrystals) are observed, the optimisation starts, and every component of the solution must be considered (buffer, salts, ions…)[59].

Crystals formation goes through different stages: supersaturation, nucleation and the effective growth (that is strictly linked to nucleation) [59]. Nucleation is the first-ordered phase transition by which protein molecules pass from a complete disordered state to an ordered one. After the formation of partially ordered (para-crystalline) intermediates, the formation of a completely ordered assembly, named as critical nuclei, occurs[59]. Therefore, nucleation is the most difficult aim to reach both theoretically and experimentally. After nucleation, the next phase is the growth of the crystal, once a stable nucleus appears in a supersaturated solution. Supersaturation is a non-equilibrium condition where some molecules exceed the solubility limit and they are no more present in solution[59]. The equilibrium is then re-established by formation and development of a solid state

(*i.e.* the crystal) [59]. To reach the supersaturation state, the properties of the undersaturated solution must be modified to reduce the ability of the medium to solubilize the protein or, on the contrary, some properties of the protein must be altered to reduce its solubility or / and to increase the attraction of the macromolecules by each other[59]. Under a practical point of view, it means to perturbate the relationship between solvent and solute in order to promote the formation of the solid state. All the components that allow the supersaturation state are named precipitants,which can be buffers that alter the pH, salts that alter the activity of the water or polymers that alter the interaction between protein and solvent. For this reason, the different combination of precipitants and their concentration are the key to force protein crystallisation[59]. For this purpose, different commercial kits for crystallisation screening are available, and this was the first approach to crystalize A075L. As the screening test is extensive, and many trials are required, the first goal is to obtain as much protein as possible. For this reason, an optimized purification method was set up, as described above in "expression and purification of recombinant A075L".

Then, the pre-crystallisation test described below, permitted to find the best concentration of both protein and buffer to obtain the correct protein precipitation avoiding amorphous formations.

The x-ray diffraction is one of the most important techniques to study the macromolecules at the atomic level. By x-ray diffraction, it is possible to obtain the electronic density of a crystal when the x-rays diffract from it. When the waves hit the crystal, they can diffract in two different ways. With the constructive interference, they give rise to a diffraction pattern, but with the destructive one, they cancel by each other. The protein crystal is a sum of constructive and destructive interferences and, in the end, a diffraction pattern will give the information about the distributions and type of atoms that compose the crystal. Then, the diffraction pattern is mathematically interpretable as described by McPherson and Gavira [59].

Pre-Crystallisation Test

The first approach to set up the crystallography screening is the PCT test (pre crystallisation test) that allows to find the best conditions in terms of precipitants and protein concentration.

The PCT kit (Hampton Research) contains 4 reagents used to evaluate protein concentration for crystallization screening. The protein concentration is evaluated with four solutions (A1, A2, B1, B2). If the sample is too concentrated itcan result in amorphous precipitate, while samples too diluted can result in clear drops. Optimizing protein concentration for the screening, is a key step in the crystallization process[60]. The PCT reagents contents are the following: **A1** 0.1 M Tris/HCl pH 8.5, 2.0 M $(NH_4)_2SO_4$, **B1** 0.1 M Tris/HCl pH 8.5, 1.0 M $(NH_4)_2SO_4$, **A2** 0.1 M Tris/Hcl pH 8.5, 0.2 M $MgCl_2$, 30% w/v Polyethylene glycol 4,000, **B2** 0.1 M Tris/HCl pH 8.5, 0.2 M $MgCl_2$, 15% w/v Polyethylene glycol 4,000[60].

96 well plate screening

According to the PCT results, 13 crystallization commercial screens for A075L were set up at 9mg/ml. in order to find the best one to have a crystal. Each plate contained 96 conditions and it was used the so called "sitting drop" vapour diffusion at 18°C described by McPherson and Gaviria, J.A., 2013 [59].

$(NH_4)_2SO_4$ (Qiagen)	JCSG+ (Molecular Dimensions)
Salt Rx1-Rx2 (Hampton Research)	Natrix (Hampton Research)
Stura Macrosol (Molecular Dimensions)	Pact (Molecular Dimensions)
Midas (Molecular Dimensions)	Proplex (Molecular Dimensions)
Morpheus (Molecular Dimensions)	Structure Screen (molecular Dimensions)
Peg Ion (Hampton Research)	Index (Hampton Research)
NR-LBD (Molecular Dimensions)	

Table 5. *Commercial screens for crystallization.* For each plate is indicated the relative company.

Each plate name is referred to the commercial kit used (from Molecular Dimension or Hampton research). Each kit differs in terms of precipitants, so using more kits for screening increases the probability to find the best nucleation condition. As described previously in this chapter, A075L was derivatize with SeMet in order to derivatize the protein, as there are no models in literature to do a molecular replacement to solve the structure.At this point, SeMet A075L was used, after the PCT test, in a concentration of 9mg/ml and with UDP-Xyl in a ratio of 1:3 protein:substrate.

48 well plate screening

After the first round of screening and the identification of the best crystallisation conditions, a new screening was started in order to improve the crystals.
In particular, the best conditions were in: **MIDAS A3**: 45% polyacrylate 2100, 0.1 M HEPES pH 6.5, **MIDAS C10**: 35% polyacrylate 2100, 0.2M NH_4SO_4, 0.1M HEPES pH 7.5, **MORPHEUS A12**: 0.06M Divalents (Mg^{2+} and Ca^{2+}), buffer system 3 (0.1M TRIS-HCl-Bicine) pH 8.9, 37% MIX (MPD-Peg1K-Peg 3350).
Three plates with 48 condition in each were set up, where each row (from well 1 to 6) contained the same precipitant conditions, but different protein ratio (1:1, 1:2, 2:1) repeated in double, and each column (from A to H) contained different concentration of the components (precipitant, buffer, salts…).
Each plate is schematized in the following tables.

	1	2	3	4	5	6
A	0,24M divalents 37,5% MIX 0,1M buffer system3 pH8,9					
B	0,12M divalents 37,5% MIX 0,1M buffer system3 pH8,9					
C	0,06M divalents 37,5% MIX 0,1M buffer system3 pH8,9					
D	0,03M divalents 37,5% MIX 0,1M buffer system3 pH8,9					
E	0M divalents 37,5% MIX 0,1M buffer system3 pH8,9					
F	0,06M divalents 18,75% MIX 0,1M buffer system3 pH8,9					
G	0,06M divalents 9,37% MIX 0,1M buffer system3 pH8,9					
H	0M divalents 4,68% MIX 0,1M buffer system3 pH8,9					
Protein:Buffer ratio	1:1	1:1	1:2	1:2	2:1	2:1

Table 6. *Morpheus A12 condition.*

	1	2	3	4	5	6
A	35% polyacrylate 2100 0,2M AmmSO4 0,1M HEPES pH7,5					
B	17,5% polyacrylate 2100 0,2M AmmSO4 0,1M HEPES pH7,5					
C	8,75% polyacrylate 2100 0,2M AmmSO4 0,1M HEPES pH7,5					
D	0% polyacrylate 2100 0,2M AmmSO4 0,1M HEPES pH7,5					
E	35% polyacrylate 2100 0M AmmSO4 0,1M HEPES pH7,5					
F	35% polyacrylate 2100 0,025M AmmSO4 0,1M HEPES pH7,5					
G	35% polyacrylate 2100 0,05M AmmSO4 0,1M HEPES pH7,5					
H	35% polyacrylate 2100 0,1M AmmSO4 0,1M HEPES pH7,5					
Protein:Buffer ratio	1:1	1:1	1:2	1:2	2:1	2:1

Table 7. *Midas C10 condition*.

		1	2	3	4	5	6
HEPES	A	45% polyacrylate 2100 0,1M buffer					
	B	22,5% polyacrylate 2100 0,1M buffer					
	C	11,25% polyacrylate 2100 0,1M buffer					
	D	0% polyacrylate 2100 0,1M buffer					
Buffer system 3	E	45% polyacrylate 2100 0,1M buffer					
	F	22,5% polyacrylate 2100 0,1M buffer					
	G	11,25% polyacrylate 2100 0,1M buffer					
	H	0% polyacrylate 2100 0,1M buffer					
Protein:Buffer ratio		1:1	1:1	1:2	1:2	2:1	2:1

Table 8. *Midas A3 conditions*. In this case rows from A to D contain HEPES buffer at pH 6.5, and from E to H Buffer system 3 at pH 8.9

For the X-ray diffraction analysis, crystals are removed from the drop and transferred for 1 minute into the same solution in which the crystal grew, supplemented with 20% glycerol for cryoprotection. Then, the crystal is flash frozen in liquid nitrogen and sent to the synchrotron. Several datasets were collected at Diamond Light Source (Didcot, UK) at beamline I04. The presence of the selenium was checked by the fluorescent scan around 12658 ev (see results section). Diffraction data were integrated and scaled using XDS[61].

2. Results

2.1 A064R

2.1.1 Sequence Analysis

A064R is one of the six PBCV-1 putative glycosyltransferases and it is reported in the CAZY database as a GT-nc (glycosyltransferase not classified)[18]. BLAST analysis on NCBI *nr* database using the full-length protein indicated that A064R is encoded by a subset of Chlorella viruses, which infect *Chlorella variabilis* and are closely related to PBCV-1. It is composed by 638 residues and at least three domains can be identified (**Figure 25**).

The **N-terminal** domain (**D1**) has been already structurally characterized, confirming its inclusion among the glycosyltransferases with type A structural fold (GT-A) [57]. Beside the subset of the Chloroviruses cited above, no significant match could be found in the viral world. Best identity with cellular organisms was found with some bacteria belonging to proteobacteria; sequence alignments among representative organism are depicted in **Figure 26.** Some highly conserved amino acids are found, corresponding to Gly 11, Asp78, Arg 202, Gly 196, and Phe 13, which are important for nucleotide-sugar substrate binding. Also, the DXD motif typical for most GT-A type enzymes is present.

The **central region** (**D2**) of the sequence did not match to any already recognized domain in the databases (**Figure 24**). BLAST analysis of the region 212- 410 revealed that it displays significant identity with portions of multidomain putative proteins encoded by *Pithovirus*, *Marseillevirus* and *Klosneuvirus*. Best hits with a good query coverage for cellular organisms were found with some members of the *Planctomycetales* orders. Interestingly, sequences with about 50% identity are found in many *Prevotella* species. Thus, the identification of A064R second domain activity will help to elucidate also the function in these bacteria, which are very important components of the human intestinal, oral and also vaginal microbiota[62].

Figure 24 reports the sequence alignment of the second domain from some representative organisms. It is important to note that no specific feature for

known GT can be recognized, suggesting that this domain may represent a new GT fold.

The last **C-terminal domain** (**D3**) was assigned to the S-adenosyl methionine dependent methyl transferase superfamily. This superfamily comprises many families responsible for the transfer of a methyl group to different acceptor substrates. BLAST analysis gave significant hits with several putative class-I methyltrasferases. Aligments are reported in **Figure 25.**

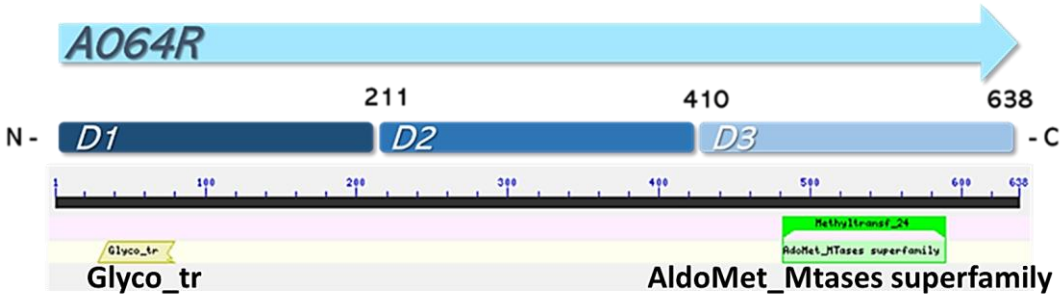


Figure 25. *A064R schematic representation.* A064R conserved domains are highlighted. It is possible to note that no conserved domains are matched for D2. D1 is recognised as GT-A domain, D3 as a methyltransferase domain.

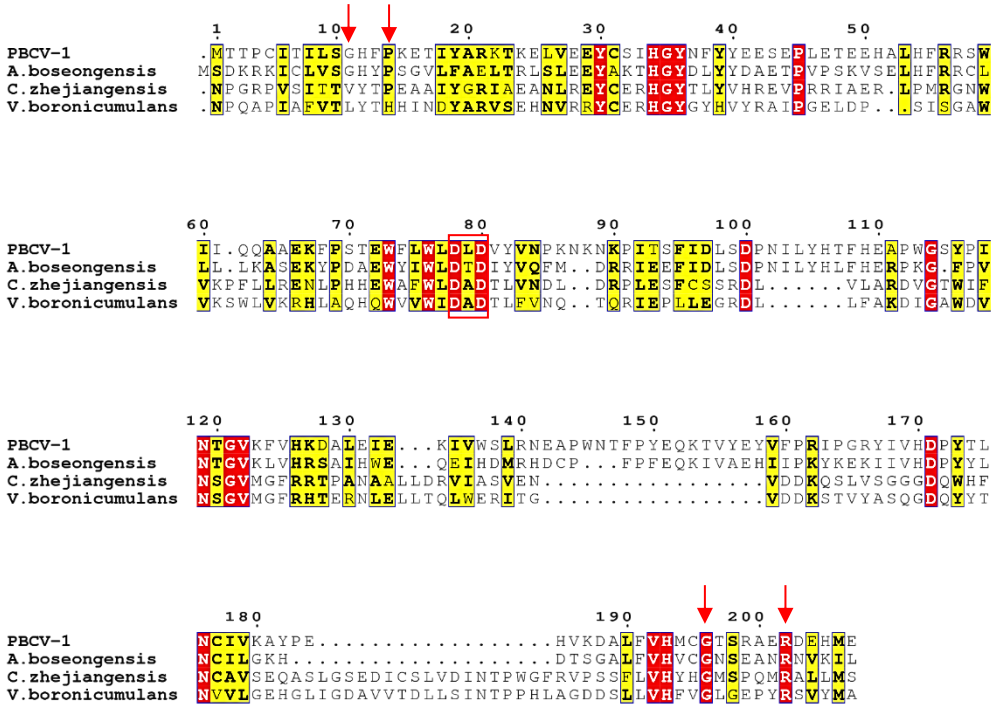


Figure 26. *Sequence multiple alignment for D1.* Conserved residues, involved in the substrate binding, are highlighted with a red row. DXD motif is highlighted in the red box. PBCV-1: AAK19297.1, *Algoriphagus boseongensis:* WP_133552711.1, *Caballeronia zhejiangensis:* KDR25164.1, *Variovorax boronicumulans:* WP_062475505.1.

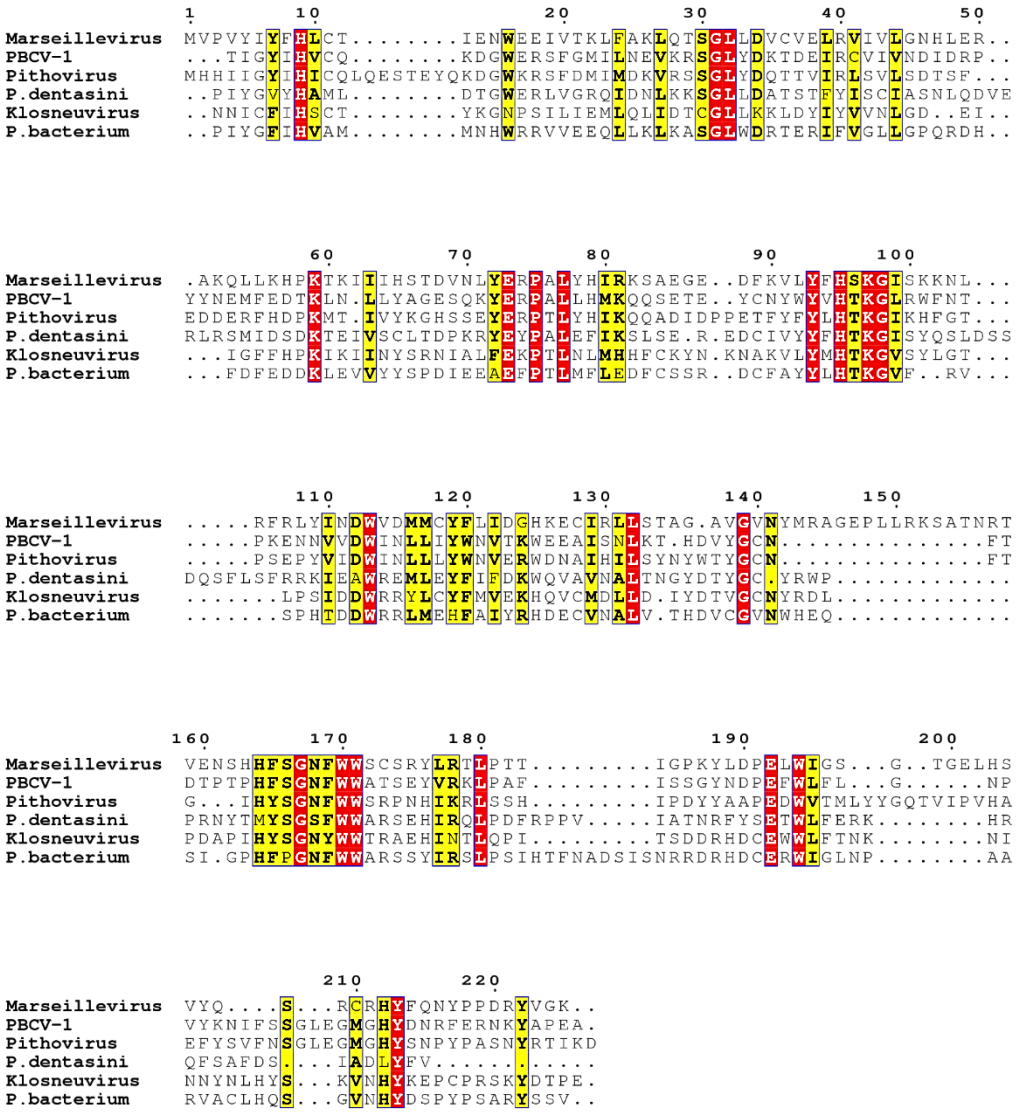


Figure 27. *Sequence multiple alignment for D2.* D2 domain does not match with any conserved domain among organisms. Best hits are sequences belonging to NCLDVs, *Planctomycetales* and *Prevotella* species.
PBCV-1: AAK19297.1*, Pithovirus:* QBK92091.1*, Marseillevirus:* QBK88047.1, *Klosneuvirus:* ARF12242.1, *Planctomycetales bacterium:* BBO34427.1*, Prevotella dentasini* BBO34427.1.

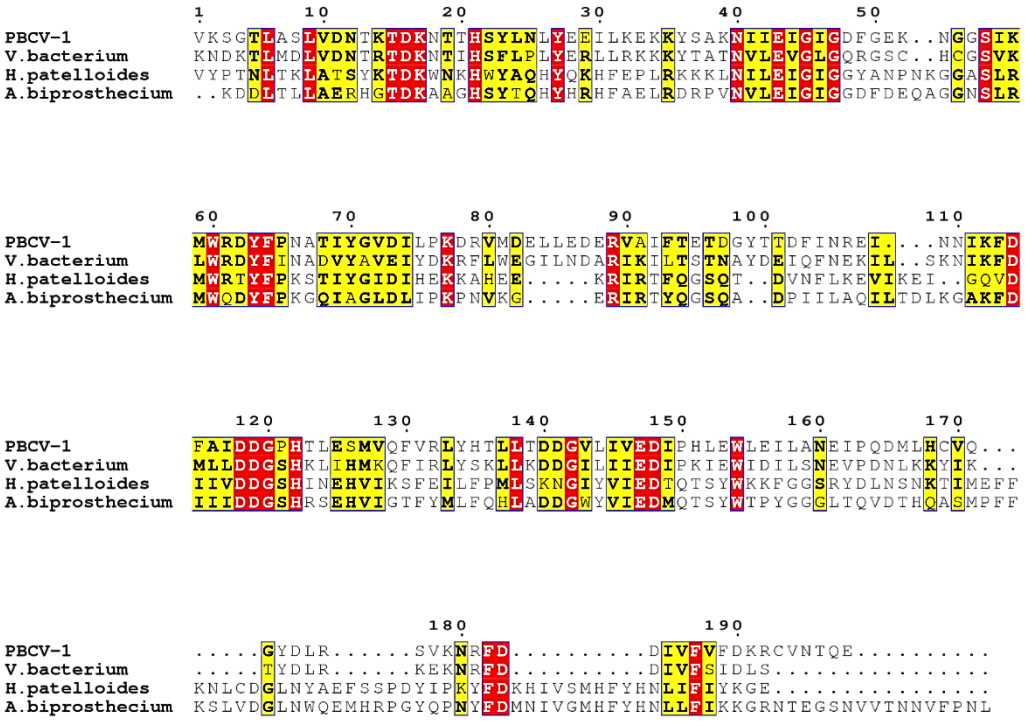


Figure 28. *Sequence multiple alignment for D3*. D3 domain matches with sequences form bacteria, identified as putative methyltransferases.
PBCV-1: AAK19297.1, *Verrucomicrobiales bacterium:* MAD25665.1, *Hyella patelloides:* WP_144864280.1, *Asticcacaulis biprosthecium:* WP_006271401.1.

2.1.2 A064R Domain Expression

On the basis of sequence alignments, A064R domains were produced as GST-fusion proteins. Since alignments were not conclusively informative to define the exact boundaries of each domains, different forms were produced, in order to identify the minimum sequence required for activity. This problem resulted particularly important for the C-terminal end of the second domain, identified as D2.

Initially, domain 1 (identified as D1, from amino acid 1 to 211, **Table 9**) and a shorter version of domain 2 (identified as D2 short, from amino acid 193 to 405, (table 7) were produced. A construct containing both domains was also prepared (D1+D2 short, from amino acid 1 to 405, table 7), in order to verify the reciprocal effects of the two regions on enzyme function. However, preliminary enzymatic tests revealed that D2 short protein was devoid of activity (see below, in the Enzymatic activity section) and that also for D1+D2 short protein only the N-terminal domain had a GT activity. For this reason, we went back to analyse the second domain sequence with more detail. Carefully inspection of the PBCV-1 variants revealed that CME6 mutant (**Figure 23**), which presents both L-rhamnose residues, but lacks methylation in the in Vp54 associated glycan[51] , presents a TC dinucleotide insertion at genome position 36,276, which leads to premature chain termination during translation of domain 3 [51]. This finding prompted us to reconsider the C-terminal boundary of D2 and to produce an extended version at the C-terminus, containing 33 extra amino acids (D2 long, from amino acid 193 to amino acid 438, **Table 9**).

Accordingly, both domains were also produced comprising this C-terminal extension (D1+D2 long, from amino acid 1 to 438, **Table 9**). Finally, the increasing availability of sequences in the databases to be used for sequence alignments suggested us to produce also a shorter version of domain 2, lacking the 17 amino acid from the N-terminal region, with the aim to identify the minimum length required for enzymatic activity (D2 long 2, from amino acid 207 to 438, **Table 9.**

PROTEIN NAME	AA	SEQUENCE DETAILS (from…to)	MW (kDa)	MW + TAG (kDa)	EXTINCTION COEFFICIENT (M^{-1} cm^{-1})	ISOELECTRIC POINT
A064R FULL LENGHT	638	1…638	74.9	75.8	150745	5.11
A064R D1	211	1…201	25.5	51	54110	5.65
A064R D1D2 short	400	1…400	47.5	74	123230	5.61
A064R D1D2 long	438	1…438	51.8	77.7	127700	5.49
A064R D2 short	212	193…405	25.15	51	70610	5.55
A064R D2 long	245	193…438	28.9	54.8	73590	5.23
A064R D2 long 2	231	207…438	27.2	53.4	73590	5.31

Table 9. *A064R domains.* The table displays the number of amino acids of each domain and the region that was cloned, the molecular weight (MW) with and without the tag used for the purification, and the parameters in terms of extinction coefficient and isoelectric point. Parameters were calculated using the ProtParam tool on the Expasy platform (https://web.expasy.org/protparam/).

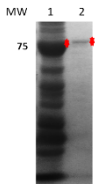
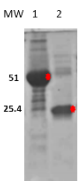
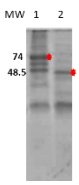
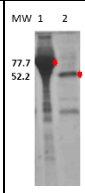
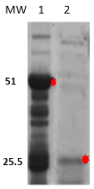
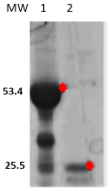
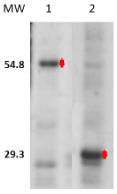
A064R full-lenght	*A064R D1*	*A064R D1D2 short*	*A064R D1D2 long*	*A064R D2 short*	*A064R D2 long2*	*A064R D2 long*
						

Table 10. *A064R domains SDS-PAGE.* For each recombinant protein the GST-fusion protein (lane 1) and the soluble protein after protease cleavage (lane 2) are shown. A064R full length, A064R D1 and A064R D1D2 long displayed high solubility, indeed, they were the proteins with the higher enzymatic activity (see following paragraph). A064R D2 short, A064R D1D2 short, A064R D2L and A064R D2L2 were more difficult to obtain in a soluble form and, as described in the following section.

All the recombinant proteins were produced using the same growth conditions and purified depending on the tag used (GST or 6xHis) (see experimental procedures section for more details). **Table 10** displays the SDS-PAGE for each recombinant protein.

A064R full-length, A064R D1, and A064R D1+D2 long are obtained as soluble recombinant proteins, with high enzymatic activity. In fact, as we demonstrate, D2 needs 33 additional amino acids to be active and, in addition, need to be produced coupled with D1 in order to maintain the correct folding: D2 long 2, even if displays the 33 amino acids at the C-terminal, displayed solubility problems.

2.1.3 A064R Domains Enzymatic Characterisation

To test activity, the reactions were performed using the proteins after proteolytic cleavage from the GST or, as alternative, with the GST-fusion proteins still bound on the GSH-sepharose beads. This latter procedure facilitates the recovery of the supernatant containing the reaction product and it was also particular useful for NMR analysis. The samples were then analysed by HPLC and by NMR; this latter analysis was performed either directly to monitor the reaction or after HPLC purification of the products.

A064R glycosyltransferase is supposed to act as a sequential UDP-L-rhamnosyl transferase; for this reason, the donor nucleotide-sugar substrate was synthesized in our laboratory, as described in the Material and Methods section. Conversely, the appropriate acceptor substrates were synthesized by Todd Lowarty laboratory at the University of Alberta, specifically octyl-β-xylose and the compound identified as Todd (**Figure 29**). Compounds Todd+1 and Todd+2 were expected to form by the sequential activity of D1 and D2 domains.

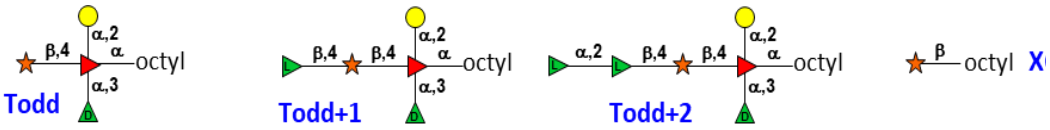


Figure 29. Acceptors used for the A064R and A064R domain enzymatic activity

The proposed activities for the D1 and D2 domains are depicted in more details in **Figure 30**.

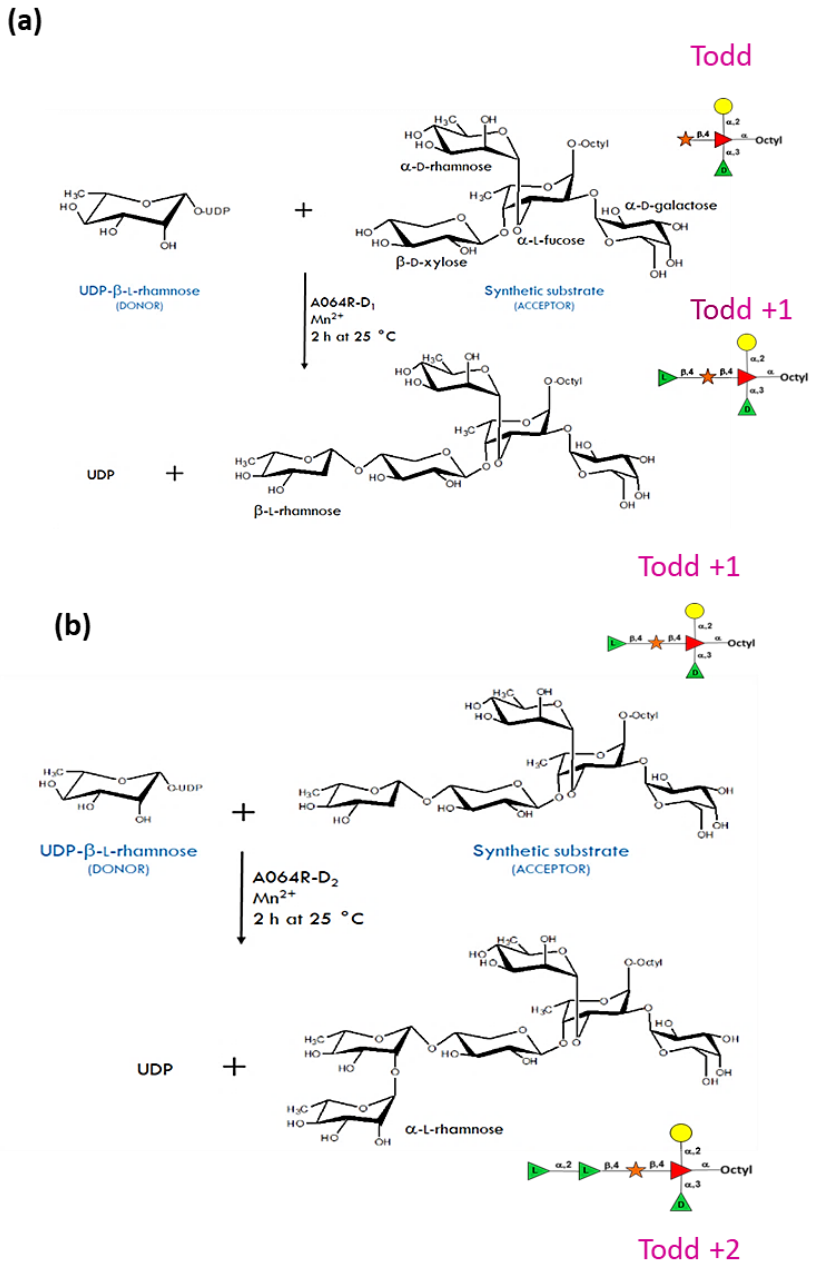


Figure 30. *Enzymatic activity proposed for D1 and D2*. Both are UDP-rhamnosyl transferases. **(a)** D1 is able to transfer the UDP-L-rhamnose on the synthetic acceptor named as "Todd", that displays the distal xylose. **(b)** At this point, D2 can attach the second rhamnose.

As described in the “Experimental procedure” section, each A064R form was tested for the enzymatic activity in the presence of the appropriate substrates. Bivalent cations were also added or omitted; EDTA was also used, to chelate residual ions that remained bound to the proteins during the purification procedures, in order to confirm the effect of cation presence. The reactions were verified by HPLC at T_0 and at different incubation times. After HPLC purification, the product was also analysed by NMR, in Prof. De Castro laboratory, to confirm its structure, and in particular the linkage type and configuration.

As expected, **A064R D1** was able to transfer the first UDP-rhamnose to the acceptor substrate “Todd” forming the compound named Todd+1 (**Figure 31**), according to the scheme in **Figure 30**, panel (a)). This reaction was performed in the presence of the bivalent cation Mn^{2+} (**Figure 31**, panel (a)). When Mn^{2+} was omitted and also EDTA was added to chelate the residual ions, no activity could be observed, indicating the need of bivalent cations for activity, as also reported for most GTs (**Figure 31**, panel (b)). **Figure 31**, panel (c) reports the NMR analysis of Todd+1 product; the first rhamnose residue is linked to xylose by a β1,4-glycosydic bond, as demonstrated for the Vp54 associated glycan, thus demonstrating that D1 is a retaining GT.

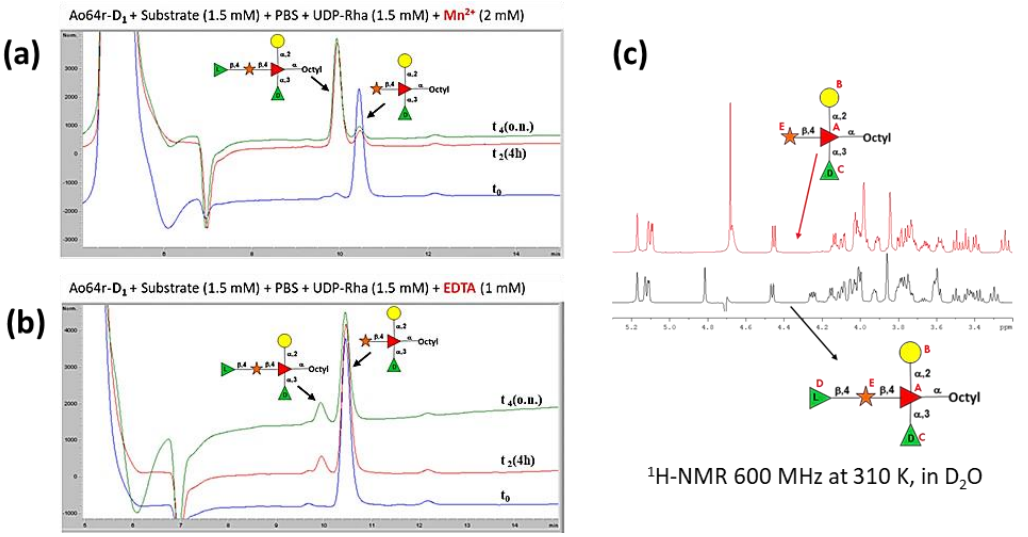


Figure 31. (a) *A064R D1 enzymatic activity*. As the HPLC chromatogram shows, the conversion of the substrate is completed after 4h with the addition of one rhamnose to the free xylose exposed by the synthetic acceptor Todd. **(b)** In addition, the incubation in presence of EDTA demonstrates that the enzyme needs cations to be active. **(c)** The NMR spectra of the purified product demonstrate that A064R D1 is a retaining glycosyltransferase, leading to the formation of a β1,4 linkage.

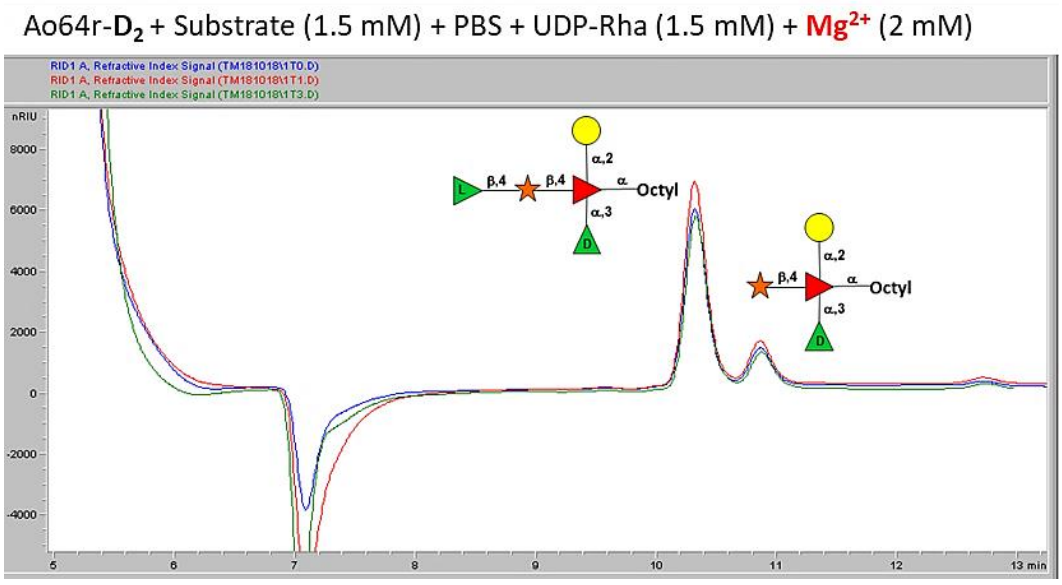


Figure 32. *A064R D2 enzymatic* activity. No product formation was detectable also after overnight incubation. (Blue line is T0, red line is T1, green line is T2).

On the contrary, it was not possible to demonstrate any activity for **A064R D2 short** domain, also in the presence of bivalent cations (either Mg^{2+} or Mn^{2+}) (**Figure 32**). To verify if D2 necessitates of D1 to acquire a correct folding and the catalytic activity, we also produced the two domains together (**D1+D2 short**). However, while it was possible to verify D1 activity, with the formation of Todd+1, D2 short did not show again any catalytic activity (data not shown).

As already indicated in the previous section, D1 and D2 were then produced together, but D2 was elongated with 33 amino acid at the C-terminus (**D1+D2 long**). This protein finally showed the two expected activities: the first responsible for the attachment of the first β1,4-linked rhamnose by D1 domain and then elongation of the first rhamnose with the second α1,2-linked rhamnose residue, catalysed by D2 domain, with the formation of Todd+2 compound (**Figure 33**). Interestingly, the bifunctional protein was able to use also octyl-xylose as substrate, indicating that the complex glycan moiety is not required for substrate recognition, but the xylose residue is the only determinant for recognition by D2 active site. This finding is particularly interesting, since it opens perspective for the use of this enzyme for glycotechnological applications.

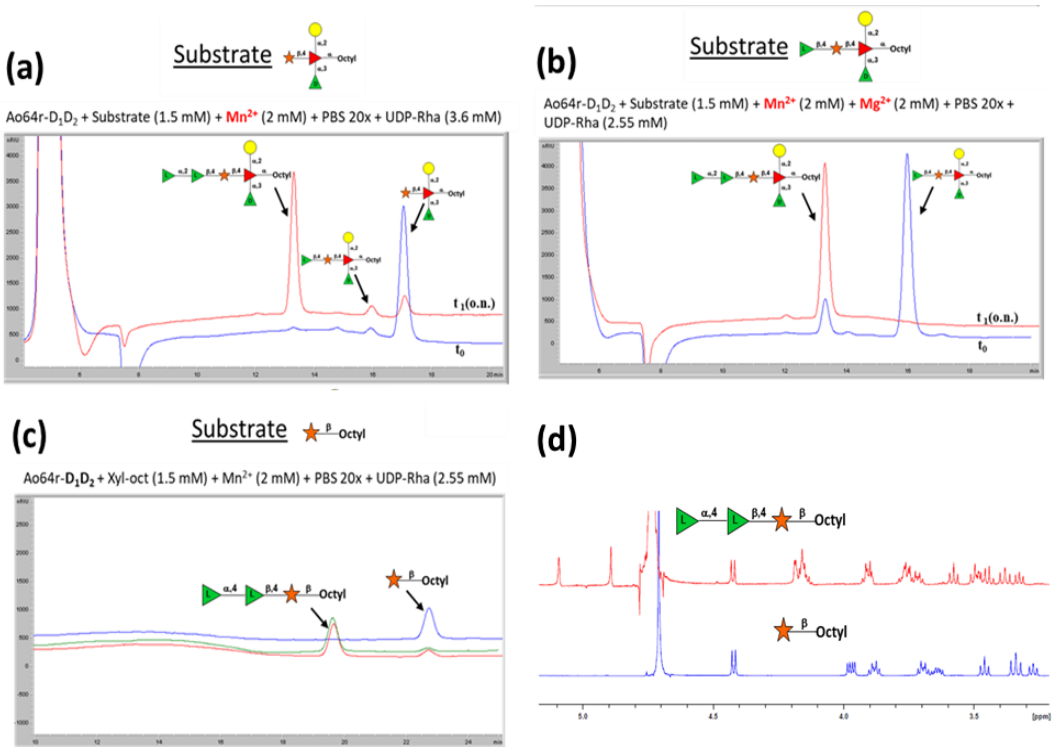


Figure 33. *A064R D1D2 long enzymatic activity.* **(a)** The reactions are carried out in the presence of "Todd", that displays only xylose; **(b)** reaction with "Todd+1" acceptor, which displays the first rhamnose. **(c)** octyl-xylose was used as alternative acceptor substrate; conversion to the product was completed after 2h (red line). **(d)** NMR analysis after the product purification of reaction **(c)** confirms the octyl-xyl-rha-rha formation.

In order to define the minimum sequence needed for enzymatic activity of the second domain, two other forms of A064R D2 were produced: A064R D2 long, which comprise amino acid 193 to 438, and D2 long2, which is 17 amino acid shorter at the N-terminus, but which has the C-terminal extension (see Table 9 and enzymatic reaction in **Figure 34**). Comparable activity and cation independence were also observed for D2 long2 protein; this indicated that the first 17 amino acids at the N-terminus of D2 long are not essential for protein structure and activity (**Figure 34**).

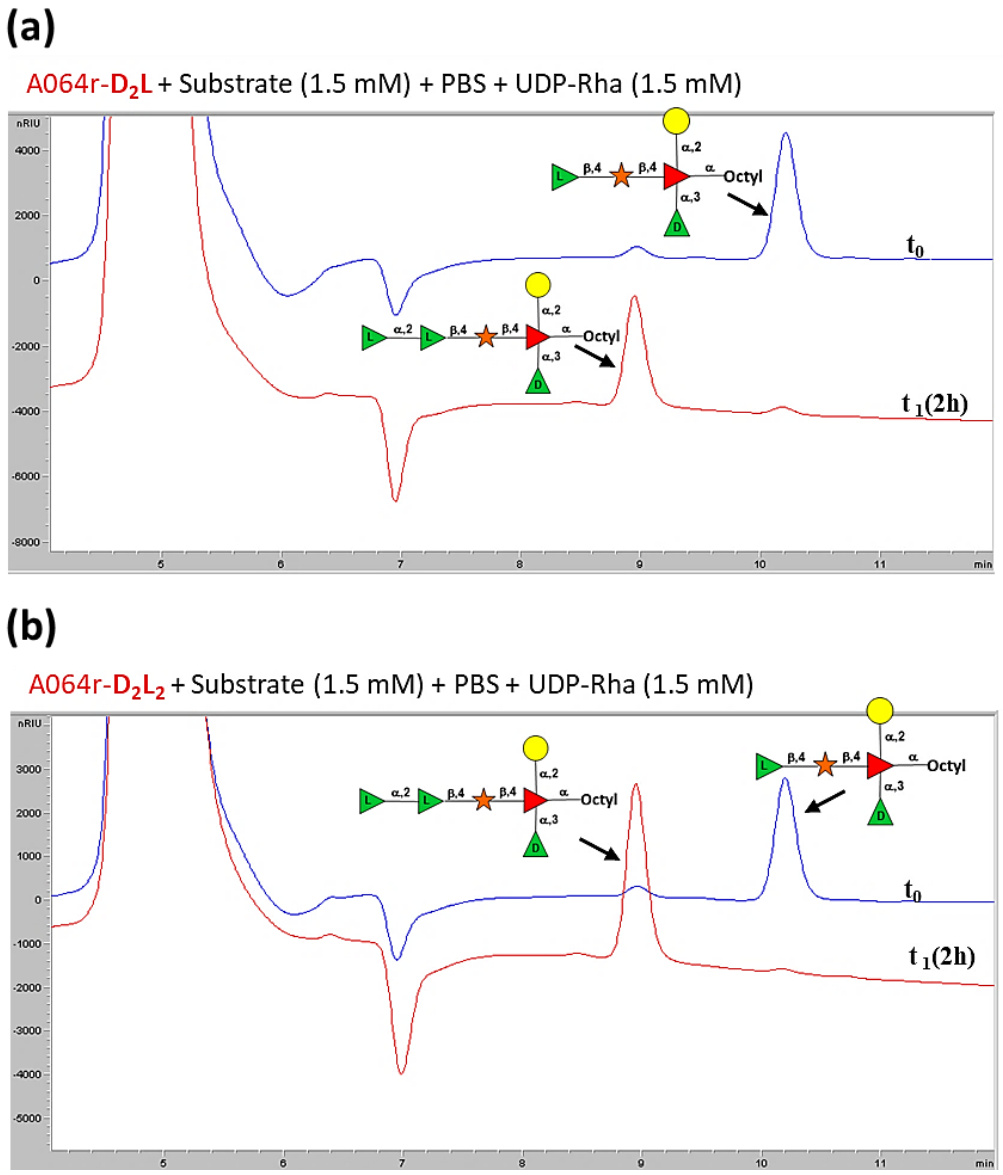


Figure 34. *A064R D2 long and A064R D2 Long 2 long enzymatic reactions. A*s is possible to appreciate from both panels, after 2h both enzymes are able to produce "Todd +2"

Enzymatic tests also revealed that D2 long does not need the addition of bivalent cations for activity. In fact, when EDTA was added to the reaction mixture, D2L maintained full activity (**Figure 35**).

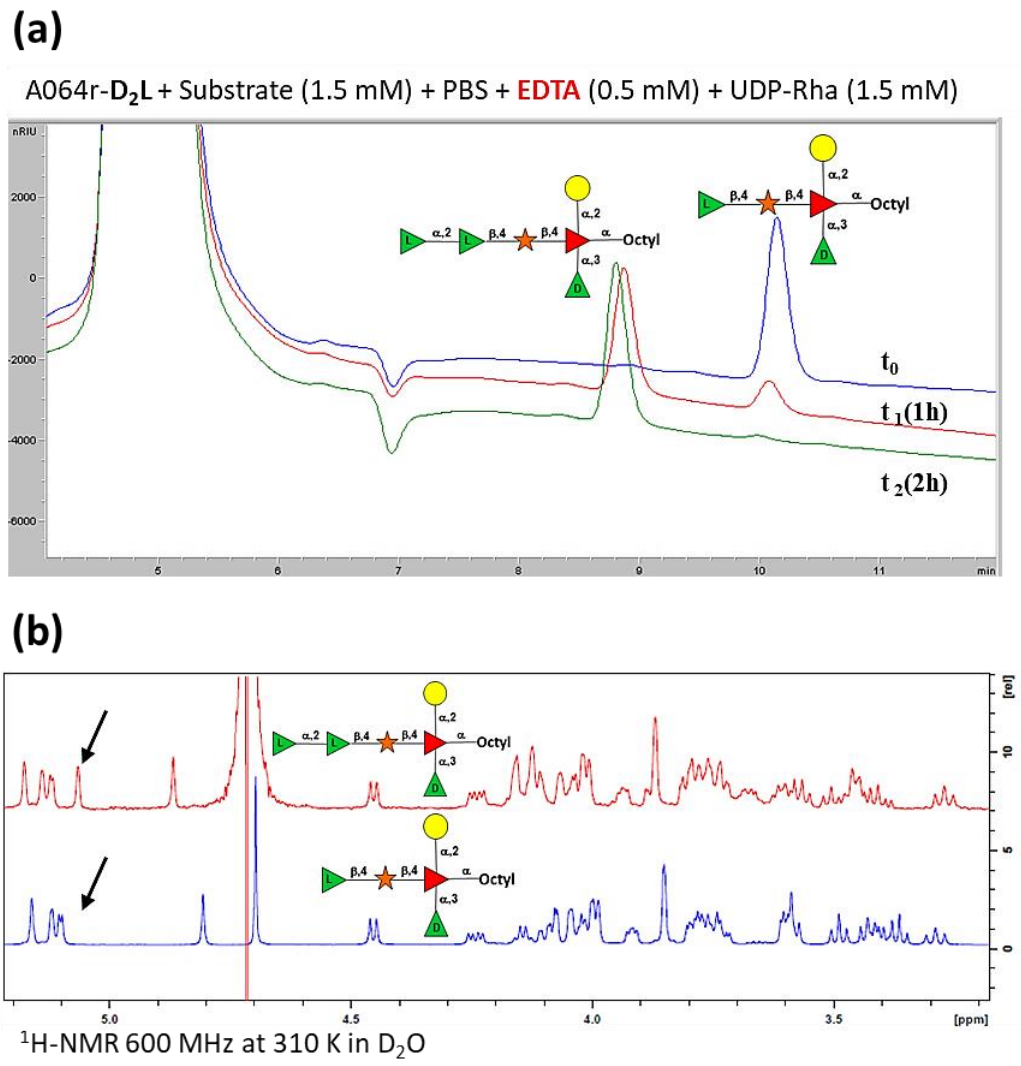


Figure 35. *A064R D2L enzymatic reaction in the presence of EDTA.* The panel **(a)** clearly demonstrate that the enzyme, in the presence of EDTA, is still active. The "Todd + 2" formation is also verified by NMR (panel **(b)**)

The cation independent activity of the second domain poses questions about its catalytic mechanism. Since D2 cannot be assigned to any known GT family, neither GT-A (which are cation-dependent), nor GT-B (which could be cation independent), it is possible that D2 represents a new, not yet recognized, type of GT fold.

D3 was not as yet produced as isolated domain; however, the possibility to produce the full-length protein allowed to test also its proposed methyltransferase activity (reaction schematized in **Figure 36**).

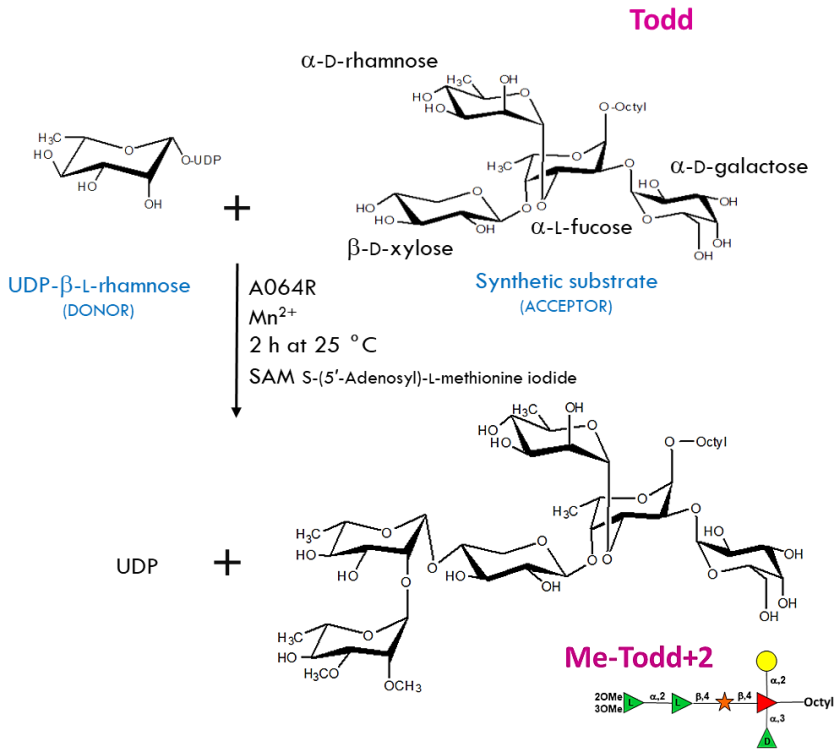


Figure 36. *Enzymatic activity proposed for D3*. D3 is supposed to methylate the distal L-rhamnose of the VP-54 glycoform.

Indeed, upon incubation of the full length protein with the nucleotide sugar donor, UDP-L-rha, the methyl donor SAM and octyl-Xyl as acceptor, it was possible to see a new specie during HPLC analysis, with a retention time higher compared to the specie with the two Rha residues (**Figure 37**). However, NMR analysis revealed the presence of only the methyl on C2 position of the distal L-rhamnose thus indicating that another methyltransferase, which still needs to be identified, may be responsible for the addition of the 3-O-methyl group, as demonstrated by NMR analysis. **Figure 38** recapitulates also the NMR spectra of the different products obtained using the A064R domains and full-length protein.

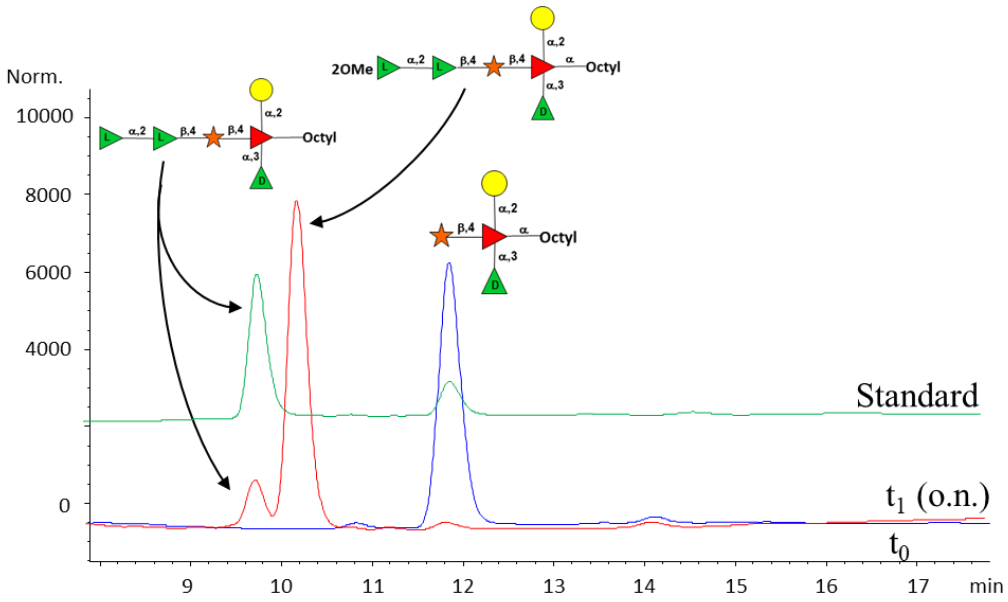


Figure 37. *A064R full length enzymatic reaction.* Using the full-length protein, D1, D2 and D3 activities are tested. After the overnight incubation, it is possible to appreciate the peak corresponding to the methylated product.

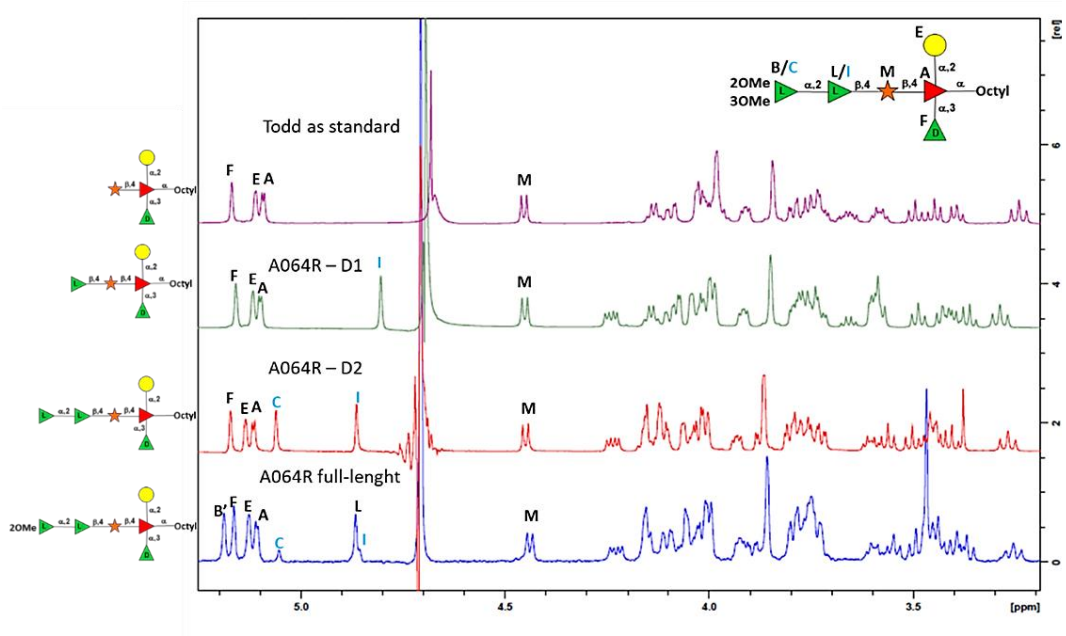


Figure 38. *Products obtained by A064R domains and full-length protein analysed by NMR.* As it is confirmed by NMR, A064R D1 catalyses the transfer of the proximal L-rhamnose to the octyl-Xyl acceptor, D2 the transfer of the distal one and D3 the capping 2-O methyl group.

2.2 A075L

2.2.1 Sequence Analysis

A075L is a 280 amino acid protein and it represents the second putative PBCV-1 GT studied so far. It is supposed to be a putative UDP-xylosyltransferase, thanks to its conservation among Chloroviruses[2] . Database searches cluster it in the exostosin family, as a GT-32 member. In other organisms exostosin proteins are responsible for the synthesis of heparan sulphate. In mammals three exostosins have been identified, named as EXT1, EXT2, EXT3, and they are ER resident glycosyltransferases[63] . The heparan sulphate repeating units do not display xylose as sugar, but they are composed by the disaccharide formed by D-glucuronic acid and N-acetylglucosamine molecules linked by *β*- (1–4) and *β*- (1–3) glycosidic bonds.

Orthologs with high identity with A075L were found in most members of the Choroviridae infecting *Chlorella variabilis* and *Chlorella heliozoae*, confirming a key role of this protein in the formation of the glycan core structure. Sequences annotated as exostosin were also found in other members of the NCLDVs group. Specifically, A075L displayed a 37% identity (E-value 7e-47) with a region of a multidomain protein from *Heterosigma akashiwo virus 1* and about 30% with *Pheocystis globosa* virus and some other members of the *Mimiviridae* family. Morevoer, conserved orthologs were identified in cellular organisms, with identities of about 30%. Alignment of A075L with representative sequences among NCLDVs is shown in **Figure 39**, and among cellular organisms in **Figure 40.**

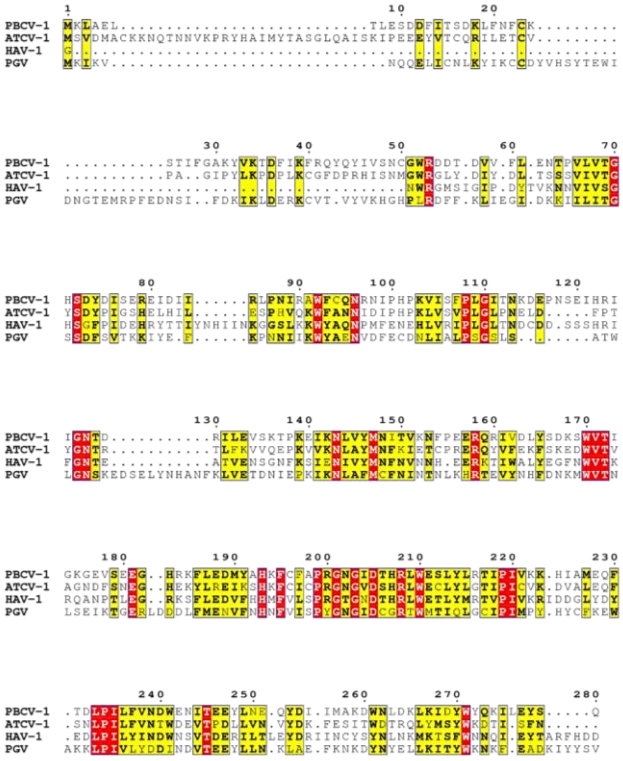


Figure 39. *A075L sequence multiple alignment*. A075L is well conserved among NCLDVs PBCV-1: NP_048423.1, ATCV-1: AGE55825.1, HAV-1: YP_009507627.1, PGV: AET72872.1.

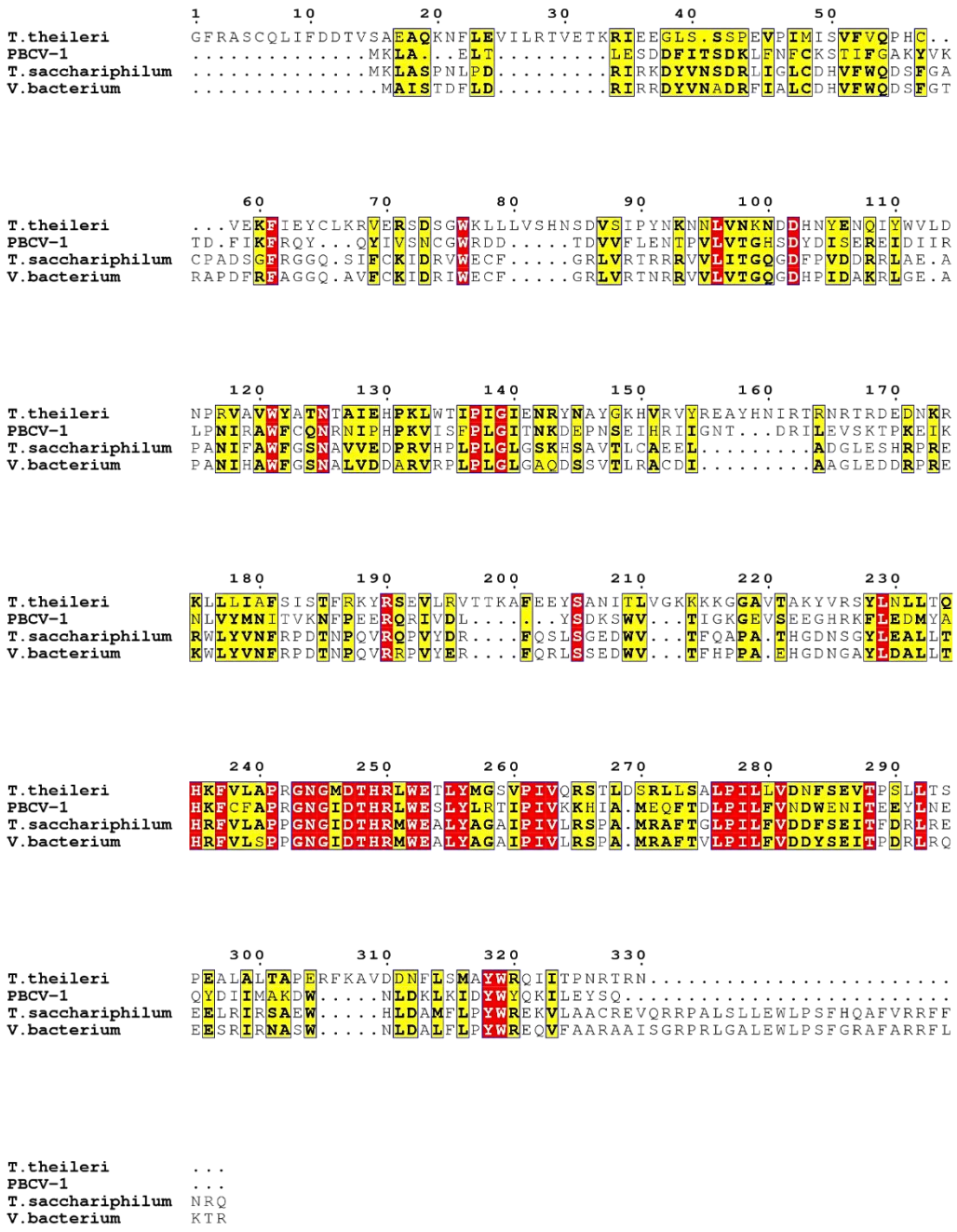


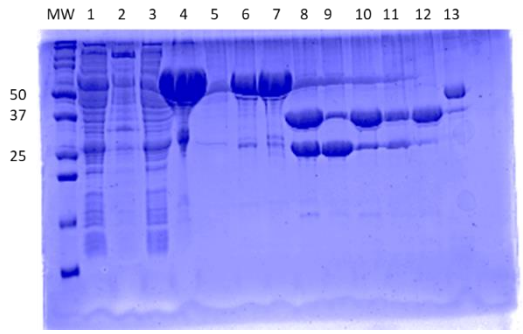
Figure 40. *A075L multiple alignment among cellular organisms.* A075L displays about 30% of identity with some bacteria and eukaryotes. The conserved domain is identified as exostosin domain that is a type of GT domain.
PBCV-1: NP_048423.1, *Terrimicrobium sacchariphilum*: WP_075077865.1, *Verrucomicrobia bacterium*: OJV11908.1, *Trypanosoma theileri*: XP_028885895.1.

2.2.2 A075L Expression

A075L purification protocol was optimized in order to obtain higher amount of protein to ensure enough quantity for the ITC experiments, the NMR experiments and crystallization screening. In this way, it was possible to perform all tests using the same batch of protein. **Figure 41**, box 1 reports the results from the initial purification steps using GSH-sepharose affinity resin and the release of A075L soluble protein from GST by proteolytic cleavage. After these initial purification steps, it was possible to obtain a quite large amount of A075L with minimal GST contamination (**Figure 41** box 1, lane 12). However, to ensure highest purity, the protein obtained after the last GSH-sepharose purification step was subjected to a further anion exchange purification: a main peak was observed, followed by a shoulder (**Figure 41**, Box 2, upper panel). All fractions were then analysed by SDS-PAGE (**Figure 41,** Box 2, lower panel). Fractions from A8 to B11, corresponding to the main peak were pooled together, concentrated to 15.5 mg/ml in 1ml, split in 50 µl aliquots, flash frozen in liquid nitrogen and stored at -80°C. The final yield was of 3,75 mg of protein par 1L of growth in LB. Fractions from B10 to B4 were considered as possible aggregate forms of the protein and appeared more contaminated by other species (**Figure 41**, Box 2, lower panel), so they were discarded.

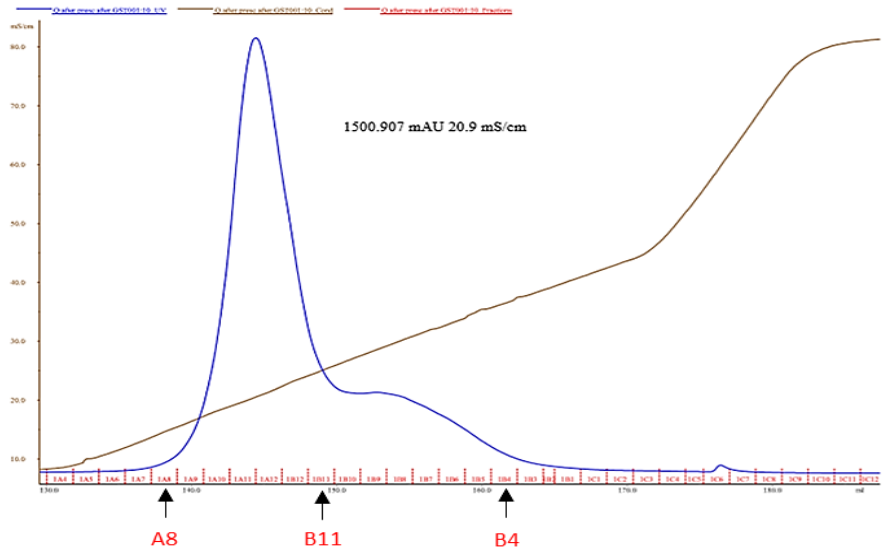
A similar protocol was used to purify the recombinant protein produced with Se-Met (**Figure 42**). In this case three peaks were observed (**Figure 42,** box 2 upper panel). Only the first main peak, corresponding to fractions A1 to A12 (**Figure 42**, box 2, lower panel) were pooled together and used for further analyses.

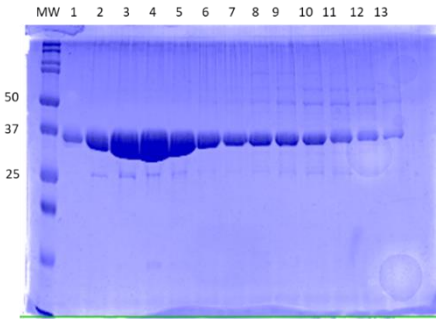
BOX 1. WT A075L purification, GSH elution, Overnight cleavage



1. Soluble protein fraction
2. Insoluble protein fraction
3. Proteins not adsorbed on the resin beads
4. Resin beads before elution
5. Wash fraction
6. Resin beads after elution
7. Elution with PBS 20 mM GSH
8. Prescission protease cleavage ON
9. Resin after GST binding 1
10. Flow through after binding 1
11. Resin after GST binding 2 (2 ul)
12. FT after binding 2
13. Prescission protease control

BOX 2. WT A075L anion exchange purification

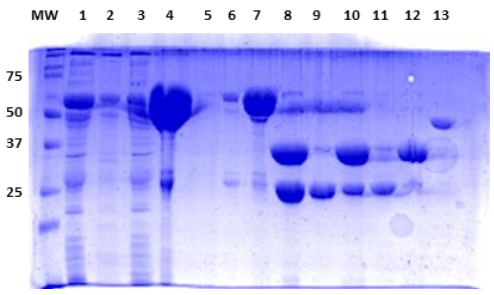




1. A8
2. A9
3. A10
4. A11
5. A12
6. B12
7. B11
8. B10
9. B9
10. B8
11. B7
12. B6
13. B5
14. B4

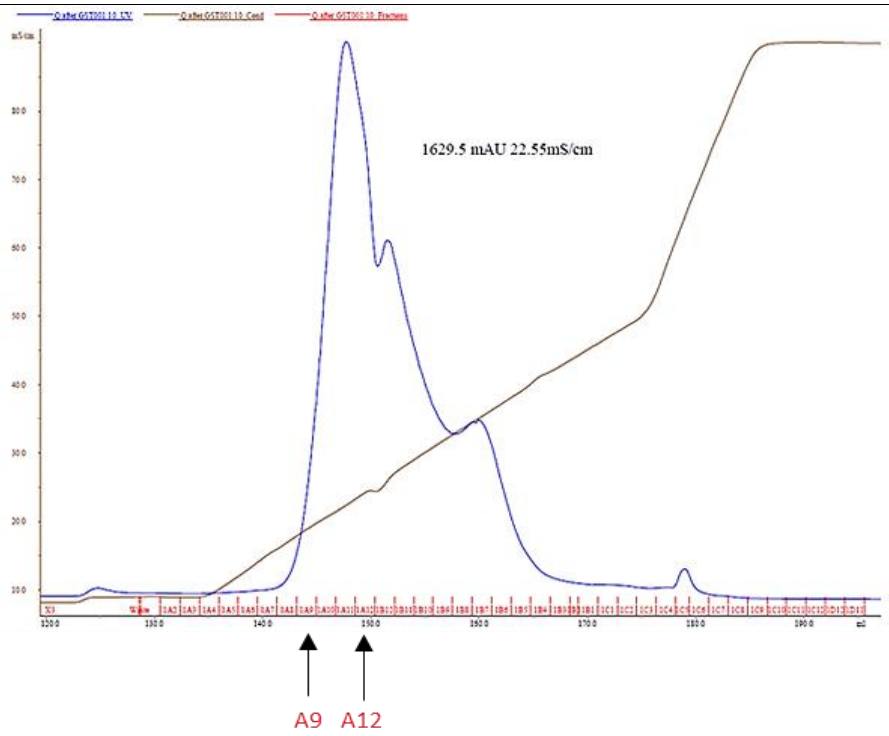
Figure 41 *A075L purification.* Box 1 displays all the purification steps verified by SDS PAGE. BOX 2 reports the anion exchange purification and the SDS-PAGE of the chromatographic fractions. Fractions A8 to B11 corresponding to "peak 1" were pooled together and concentrated. Fractions from B10 to B4, even if contained good amounts of A075L, were discarded.

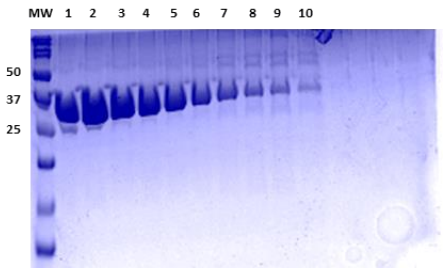
BOX 1. SeMet A075L purification, GSH elution, Overnight cleavage



1. Soluble protein fraction
2. Insoluble protein fraction
3. Proteins not adsorbed on the resin beads
4. Resin beads before elution
5. Wash fraction
6. Resin beads after elution
7. Elution with PBS 20 mM GSH
8. Prescission protease cleavage ON
9. Resin after GST binding 1
10. Flow through after binding 1
11. Resin after GST binding 2 (2 ul)
12. FT after binding 2
13. Prescission protease control

BOX 2. SeMet A075L anion exchange purification





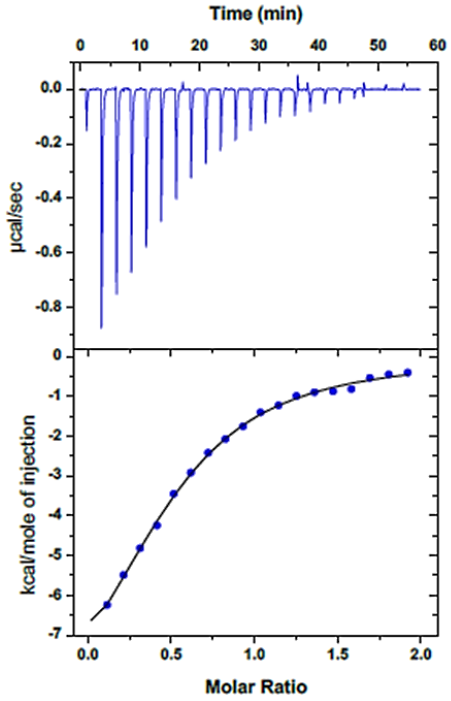
1. A9
2. A10
3. A11
4. B12
5. B12
6. B11
7. B9
8. B8
9. B7
10. B6

Figure 42. *SeMet A075L purification*. Box 1 displays all the purification steps verified by SDS PAGE. BOX 2 displays the anion exchange purification and the positive fractions verified by SDS PAGE. Fractions A9 to A12 corresponding to "peak 1" were pooled together and concentrated.

2.2.3 A075L Characterisation

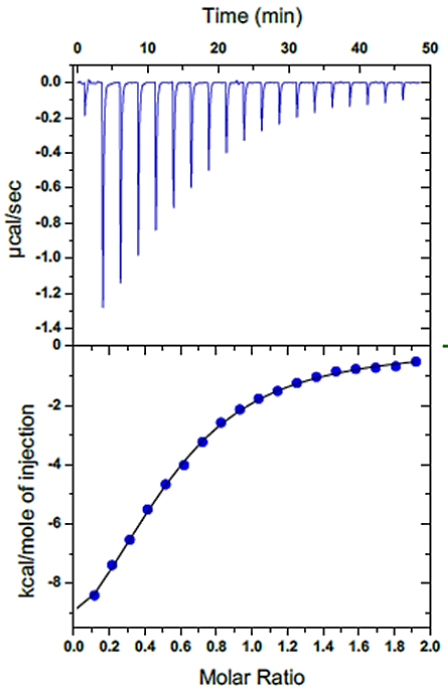
Isotermal Titration Calorimetry (ITC)

ITC was used to address the dissociation constant (K_d) of A075L with UPD-xylose. The thermodynamic parameters are shown in **Figure 43** and **Figure 44**. The results show similar K_d of A075L for UDP-xylose in presence of Mg^{2+} (28.7 µM) or Mn^{2+} (23.3 µM). The stoichiometry (N = 0.5) means that one molecule of UDP-xylose could bind to two molecules of A075L, i.e. one UDP-xylose bound to a dimer of the protein. All the experiments are repeated three times.



	UDP-Xylose → AO75L 1mM 100uM
K_d (M)	$28.7X10^{-6} \pm 4.04\ X10^{-6}$
$K_a(M^{-1})$	$42.9X10^{3} \pm 4.9\ X10^{3}$
N	$0.507 \pm 1.7x10^{-2}$
ΔH	-10.5 ±0.85
ΔG	-6.2
-TΔS	4.28

Figure 43. *ITC of A075L vs. UDP-xylose in presence of $MgCl_2$*. The top graph represents the differential heat released during the titration of 1 mM UDP-xylose into 100 µM A075L. The bottom graph represents the fitted binding isotherms. Thermodynamic parameters of A075L interaction with UDP-xylose in presence of Mg^{2+} are shown.



	UDP-Xylose → AO75L 1mM 100uM
K_d (M)	$23.3X10^{-6} \pm 1.99\ X10^{-6}$
$K_a(M^{-1})$	$34.8X10^{3} \pm 3.6\ X10^{3}$
N	$0.518 \pm 9.4x10^{-3}$
ΔH	-12.9 ± 0.558
ΔG	-6.32
-TΔS	6.60

Figure 44. *ITC of A075L vs UDP-xylose in presence of $MnCl_2$*. The top graph represents the differential heat released during the titration of 1 mM UDP-xylose into 100uM A075L. The bottom graph represents the fitted binding isotherms. Thermodynamic parameters of A075L interaction with UDP-xylose in presence of Mn^{2+} are shown.

Nuclear Magnetic Resonance (NMR)

NMR experiments were performed in order to confirm the binding of A075L to UDP-xylose. UDP-xylose was incubated with the purified A075L protein directly in the NMR tube and the reaction was monitored with time.

Results shown in **Figure 45** indicate that a progressive hydrolysis occurs, with a formation of free xylose. Upon addition of the enzyme to UDP-xylose, the quick formation of β-xylose could be appreciated in NMR spectrum, in the red trace, corresponding to time 0, is indicated by an asterisk. Peak height increased at 15 minutes, with the concomitant appearance of the α anomeric specie (green trace, **Figure 45**). The initial formation of the β-anomer could suggest an inverting reaction mechanism, as it can be expected for the xylosyltransferase activity which use UDP-α-xylose as donor.

Since the hydrolysis reaction proceeded quite slowly, it was also tested in the presence of ions (Mg^{2+} or Mn^{2+}), but a significant increase of the reaction speed was not appreciable for both ions also at long incubation time (**Figure 46**). According to this information, the enzyme probably needs to bind also the acceptor substrate to promote the reaction. The effects of bivalent ions still remain controversial and further analysis are required to solve this issue.

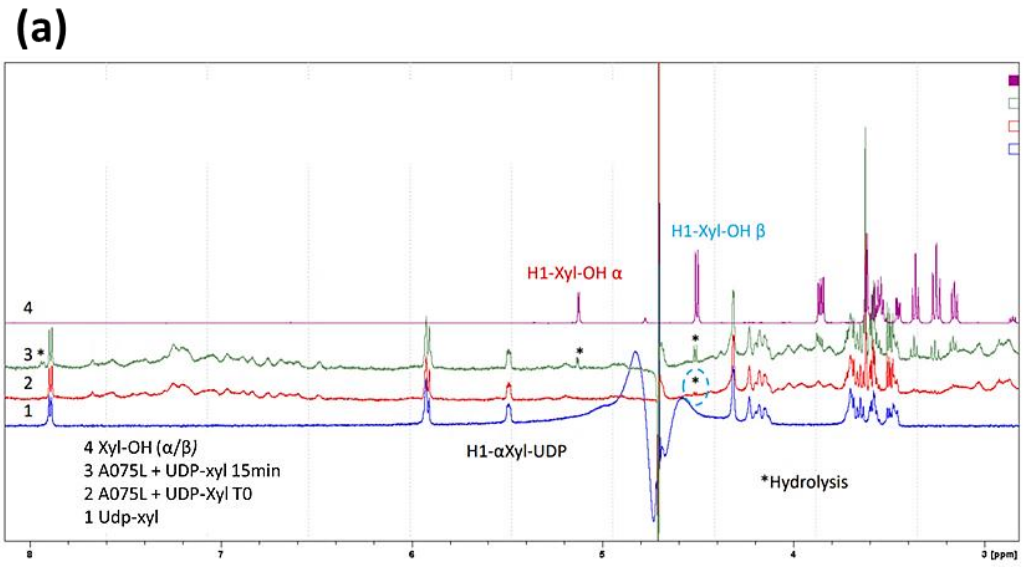


Figure 45. *A075L is an inverting glycosyltransferase.* The results show that A075L hydrolyzes UDP-α-xylose possibly with an inversion of the anomeric configuration of the xylose. Although βXyl is the major anomer in equilibrium for free xylose (Approx 70:30), it is detected immediately at the beginning of the incubation (spectrum 2)**.**

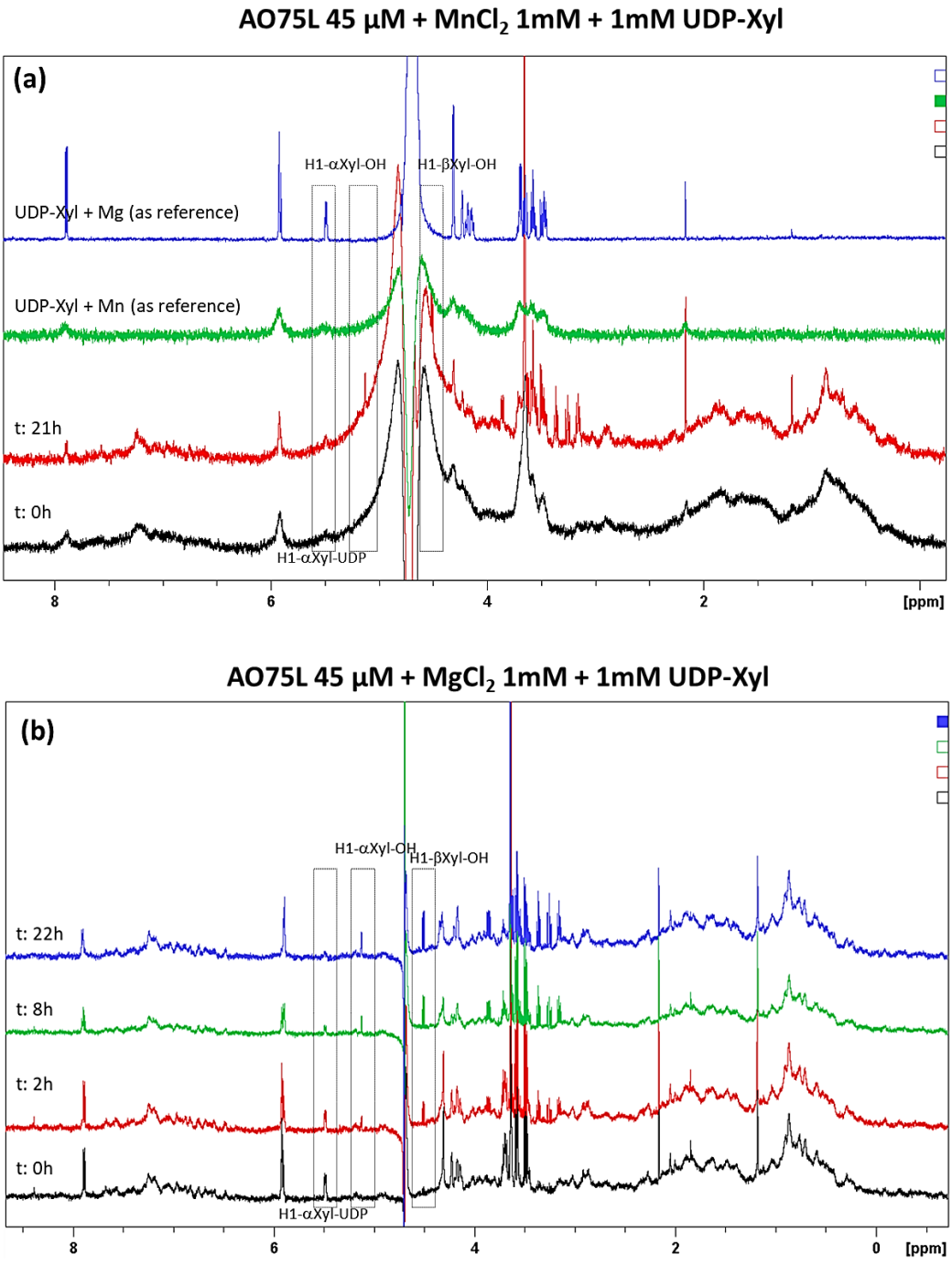


Figure 46. *A075L and UDP-xylose in the presence of divalent ions.* Panel **(a)** shows the reaction in the presence of Mn^{2+}. because of Mn^{2+} is paramagnetic, the reaction is difficult to follow, because of the background. Panel **(b)** shows the reaction in the presence of Mg^{2+}.

Enzymatic reaction of A075L

As the ITC and NMR experiments demonstrate, A075L binds UDP-xylose; however, the hydrolysis reaction was slow, indicating the need of the acceptor to promote the enzymatic activity. Thus, enzymatic tests were set up to demonstrate its catalytic activity.

Free fucose and octyl-fucose were initially used as acceptor substrates, in the presence of bivalent cations; indeed, previous experiments with A064R D1 domain showed that the presence of a complex oligosaccharide structure was not necessary to promote the enzymatic activity. Reactions were monitored by NMR, but no product formation was observed. This finding suggested the enzymes requires a more complex structure, i.e the core region of Vp54-associated glycan, for recognition as substrate. Presently this structure needs to be chemically synthetized, or as alternative, a glycopeptide displaying the truncated glycan could be purified from E11 or E1L3 mutants (see **Figure 23** in introduction section) as described by Speciale et al [51].

Indeed, preliminary results were obtained using a small amount of the purified E11 glycopeptide incubated with UDP-xylose and they showed that A075L is able to transfer the xylose unit to the acceptor, thus definitively confirming its enzymatic activity and indicating that a more complex glycan structure is needed for recognition of the acceptor substrate by the enzyme active site (**Figure 47**). Unfortunately, the amount of the glycopeptide that can be purified from E11 mutant is very low, due to virus instability and poor recovery after the end of the infectious cycle. For this reason, a more detailed enzymatic characterization was not feasible. However, chemical synthesis of suitable glycan acceptors is currently undergoing.

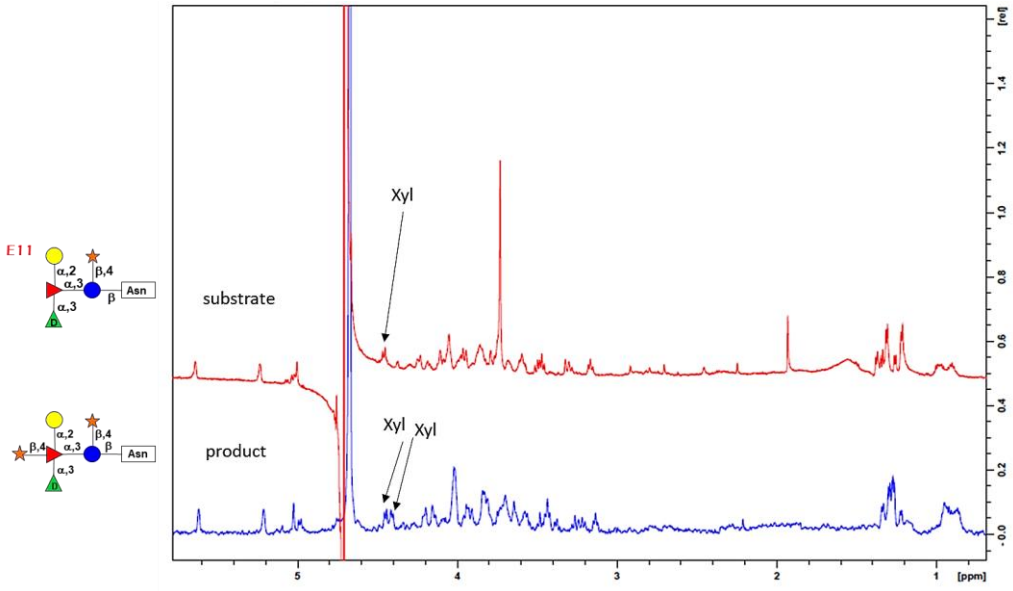


Figure 47. *A075L enzymatic reaction*. A075L was incubated overnight in the presence of the acceptor, directly purified from de mutant variant E11, and with both cations Mg^{2+} and Mn^{2+}. This preliminary data, coupled with the demonstration that A075L is able to bind the UDP-xyl, suggest that A075L is a UDP-xylosyltransferase.

2.2.4 A075L Structural Characterization

Crystallization procedure

After protein purification and concentration, the pre crystallization test (PCT test) was set up to find the best condition for crystallization in terms of salts, polymers and protein concentration. In fact, samples too concentrated can result in amorphous precipitate, while samples too diluted can result in clear drops. Precipitate and clear drops are typical crystallization results during screening, due to reagent conditions that do not promote crystallization and are formed in every crystallization screen. However, by optimizing protein concentration, it is possible to enhance chances for crystallization. For this reason, PCT can minimize or prevent situations where a screen results in an overabundance of precipitate or clear drops [60]. According to the PCT test results, the best concentration of protein was around 9 mg/ml.

This first screening is necessary to validate as much combinations as possible, in order to identify the ones where the protein concentration and the solution components can create the best condition to allow protein crystallization. In fact, each well contains different conditions of salts, pH, buffer and polymers, all of them defined as precipitants. For the screening, 13 plates (prepared using WT A075L and SeMet A075L) with 96 different condition in each were set up, using the available commercial kits. After a week some nucleation was observed. According to this result, a further screening with 48 wells plates was started. The 48 well plates allowed also to use higher amounts of protein, in order to obtain bigger crystals.

First screening was performed using the underivatized A075L protein (WT A075L), in order to find the best conditions. Finally, after identification of the best conditions, A075L was labelled with Selenomethionine(SeMet A075L), since no crystallographic structures of homologs proteins are available for molecular replacement[64].

Figure 48 shows the MALDI-TOF analysis of SeMet A075L, which shows an increase of molecular weight of the protein from 33.597 to 33.806. That indicates the

incorporation of 5 residues of SeMet (149.21), as expected for the sequence of the protein (**Figure 49**).

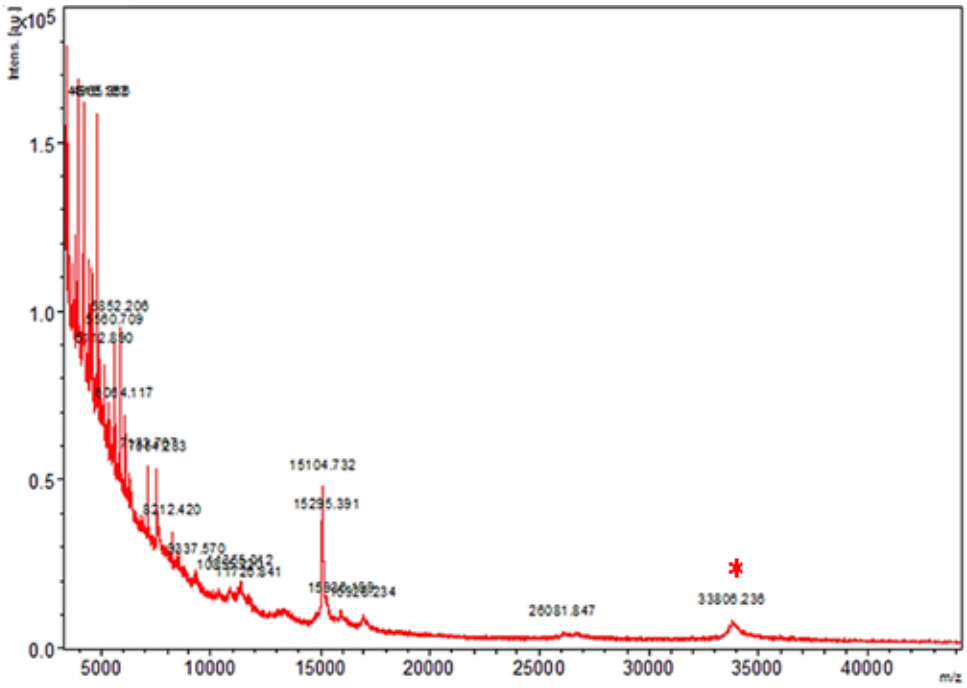


Figure 48. *SeMet A075L MALDI-TOF spectrum.* SeMet A075L was analysed by mass spectrometry in order to confirm the presence of selenomethionine. As the WT protein molecular weight is around 33,5 kDa, it can be assumed that the increase of the molecular weight is due to SeMet incorporation.

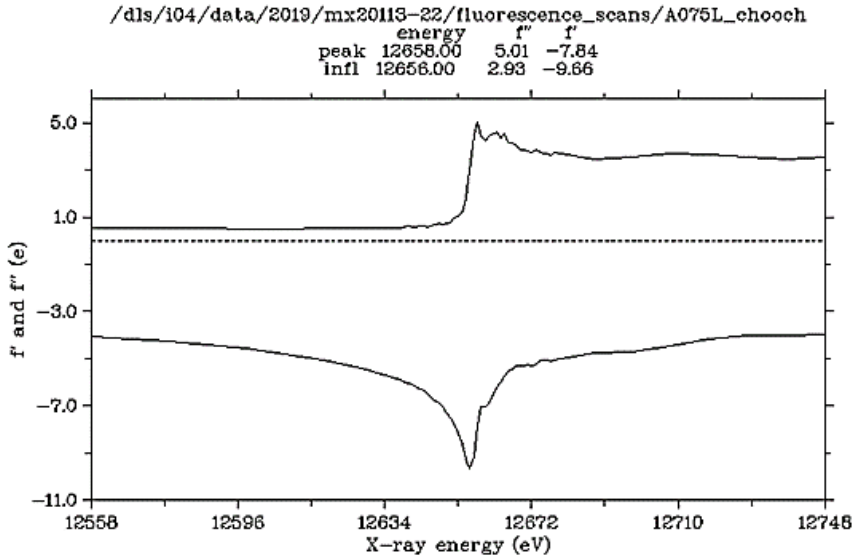
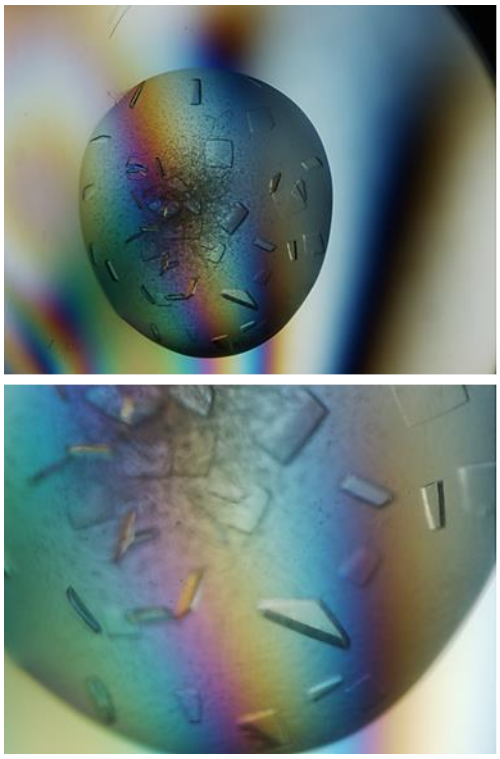


Figure 49. *Fluorescence scan of one SelMet Crystal at the selenium edge.*

The conditions that allowed the formation of the best crystals were the ones in the Morpheus screening (see **Table 6.** Morpheus A12 condition. This 48 well plate displayed a lot of crystals that were removed from the drop and flash frozen in liquid nitrogen, then stored as described in experimental procedure section, in order to be sent for the X-ray diffraction. One SelMet A075L dataset diffracted at 3.1 Å and the space group was determined to be P1, phasing of the structure is still in progress.

Figure 50 depicts a representative drop of crystals used for X-ray analysis and the corresponding data collection statistics. The data processing for determining the 3D structure are still in progress.



	A075L
Data collection	
Space group	P1
Cell dimensions□□ *a*, *b*, *c* (Å)	87.40 90.60 91.40
□□□□ *α*, *β*, *γ* (°)	90.100 108.600 109.500
Resolution (Å)	49.5-3.0 (3.17-3.00)
R_{meas}	0.08 (0.24)
$CC_{1/2}$ (%)	97.0 (45.5)
$I/\sigma I$	3.5 (1.0)
Completeness (%)	90.2 (87.0)
Redundancy	1.8 (1.7)

*Highest resolution shell is shown in parenthesis.

Figure 50. *SeMet A075L Crystals and data collection statistics.* SeMet A075L crystal was analysed by X-ray diffraction. The resolution was of 3.1 Å and the space group was P1.

3. Discussion

The interest in giant viruses has grown over the past 30 years, from *Acantamoeba mimivirus*[1] discovery in 1992 by Timothy Robotham, a microbiologist at Leeds Public Health Laboratory. Since then, the research in this topic has expanded very fast and many information obtained by scientist all over the world are available. First, it is now well established that giant viruses are widely diffused in the environment, infecting eukaryotic organisms such as plants, algae and animals. Especially, metagenomic studies in the oceans have established that giant viruses are probably the most represented viruses after bacteriophages in the sea water, influencing directly or indirectly many biogeochemical and ecological processes, including biodiversity and genetic transfer, CO_2 fixation and carbonate sinking, nutrient cycling and algal bloom control[7,48]. Secondly, according to phylogenetic studies, it has been also demonstrated that they share a common ancestry[65], which classifies them in a monophyletic group, indicated as Nucleo-Cytoplasmic Large DNA Viruses (NCLDVs). At the same time, genes from bacterial, archaeal or eukaryotic origins are commonly found in the same virus, suggesting also that extensive horizontal gene transfer occurs. At this purpose, viruses could represent a kind of "melting pot" for different genomes[26].

The main characteristic that has emerged from the studies on giant viruses is the huge particle dimension, which is necessary to contain the large genome. In fact, they possess a double stranded DNA genome ranging from 0.3 to up to 1.2 Mbp, this being a size comparable to that observed for cellular organisms. Additionally, genome analysis of sequenced viruses has revealed the presence of genes that are not usually found in viruses, but that are typical of eukaryotic organisms and are able to confer partial independence from the cell host. Indeed, it was demonstrated that they possess the replication machinery, proteases, part of the translation system, enzymes involved in redox reactions and the enzymes involved in glycosylation and modification of glycans [39]. NCLDVs also encode for a large set of genes exclusively found in this kind of viruses, which functions are presently unknown. For instance, of the 900 putative genes of Mimivirus, only 25% of them have a homologue in databases[7].

As NCLDVs are cellular parasites, but that maintain partial independence from the host mechanisms, a new definition of "virus" is going to be defined. As suggested by Forterre and colleagues, the question "are viruses alive?" probably will need a new answer[24]. To date, viruses were always considered not alive organisms, which depend totally from the infected hosts. Now, there are on the contrary evidences and demonstration that giant viruses display uncommon characteristics for a virus, giving them the new definition as "giruses", comprising the information of big dimension (giant) in term of viral particle and large genome[1], and proposing, more recently, the new order of "Megavirales", as suggested by Claverie et al.[44]

In the present work, we characterized two glycosyltransferases identified in *Paramecium Bursaria Chlorella virus-1* genome. Previous PBCV-1 genome analyses suggested the presence of an autonomous glycosylation system in this virus. This hypothesis was confirmed by the characterisation of the glycan structure exposed by the main capsid glycoprotein Vp-54. In fact, the glycoforms identified and solved by De Castro et al[50] show some peculiarities that presently pose them as unique in all kingdom of life, specifically the β-Glc linked to Asn, which is not found in the typical consensus sequence for N-linked glycosylation, and finally the complexity of the glycan structure. Thus, solving of the glycoform structure has confirmed that PBCV-1 encode for its own glycosyltransferases. At least 6 (A111/114R, A546L, A075L, A064R, A071R A219/222/226R, A473L) have been identified.

A064R is encoded only by PBCV-1 and few closely related Chloroviruses and it is not found in any other NCLDVs[39]. Discovery of PBCV-1 spontaneous glycosylation mutants provided important information to understand how the glycoforms are synthetized. The mutants analysed, in fact, showed Vp-54-associated aberrant glycoforms and most mutations were found in A064R gene[51], further indicating its glycosyltransferase activity. The reduced stability of the virus and the lower yield of the mutant viral particles observed after the end of the infection also suggested an important role for A064R in virus life cycle[51].

A064R is a multidomain enzyme and, thanks to the spontaneous mutant studied and thanks to database search, we identified in the protein two glycosyltransferases domains (D1 and D2) and one methyltransferase domain

(D3). As this work demonstrates, D1 and D2 are rhamnosyltransferases that attach the two distal rhamnose of the glycoform, confirming the bioinformatic analysis. D3, as a putative methyltransferase, is identified thanks to the identification of a mutant that lacks on the methyl groups on the last distal rhamnose, coupled with a mutation in the D3 gene region. Preliminary experiments using the full length protein have confirmed D3 enzymatic activity, but deeper characterization is still undergoing[51].

It is important to note that often the giant virus glycogenes are clustered in multidomain enzymes that ensure a metabolic channeling of substrates. For A064R, the three enzymatic activity are sequentially arranged, and the second domain is able to transfer the nucleotide-sugar on the acceptor only after the addition of the previous sugar by the first domain. This mechanism is particularly important to increase the rate of glycan formation in the viral factories and it represents a parallel to what happens in cellular organisms. Indeed, in particular in eukaryotes the sequential and coordinated activity of the GTs is ensured by the fact that they are mostly membrane bound and they are localized in specific area of the secretory pathway, where they can also form hetero oligomers. Other examples for a multidomain GT in PBCV-1 is represented by A111/114, which is currently under study.

A second interesting feature of the rhamnosyltransferase is the broad acceptor specificity. In fact, D1 is able to catalyse the addition of the proximal L-Rha to the simple xylose monosaccharide or to a lipid linked xylose, indicating that the complex core glycan is not essential for recognition. Moreover, also D2 attaches the second L-rha to first one, without needing a more complex olisaccharide for interaction with the active site. This open perspective for the use of these two domains, which are soluble and produced in good amounts, also for glycotechnological application. Further studies are needed to better understand their substrate specificity and catalytic properties.

Particularly interesting is A064R D2 domain. In fact, it does not match with any already identified domain in database search in any superkindoms of life, suggesting that it may represent a new GT fold. It is also a retaining, cation independent, GT. Since homologous sequences are present in several bacterial

species, in particular in Prevotella, the characterization of this new GT type will open the way to identify new still unknown glycogenes.

The second enzyme that has been characterised in this work is A075L. A075L is not exclusive of PBCV-1, but it is present among most Chloroviruses, a subfamily of Phycodnaviruses where PBCV-1 is the viral model[2]. A075L was indicated as the UDP-xylosiltransferase, responsible of the transfer of the distal xylose on the core fucose, but it does not have significant homology with other xylosyltransferases already characterized in the literature. However, the bioinformatic analysis assigns it an exostosin domain, that is a inverting GT involved in heparan sulphate synthesys in both prokaryots and eukaryots[18]. This finding is quite surprising since exostosins are processive enzymes, which lead to the formation of long polysaccharide chains and are supposed to remain bound to the glycan substrate during monosaccharide addition.

Both NMR and ITC experiments on A075L demonstrated that the enzyme binds the substrate UDP-xylose. Incubation of the enzyme alone with the nucleotide sugar donor in the absence of the acceptor causes hydrolysis, which seems to proceed following an inverting mechanism. In addition, the K_d from the ITC analysis suggested that one molecule of UDP-xylose can bind two molecules of the protein, suggesting that the enzyme works as a dimer. However, no informations are currently available on A075L quaternary structure. Since a good diffraction spectrum was obtained from the last X-ray crystallography experiments, A075L structure is going to be solved and this will provide an explanation about this finding.

Testing the enzymatic activity of A075L proved to be more complicated, compared to A064R. In fact, while A064R first domain displayed a broader substrate specificity, with rhamnose monosaccharide being the only determinant for recognition by the active site, A075L did not show any catalytic activity on free fucose or lipid -immobilized fucose. Conversely, it was able to transfer a xylose unit on the E11 mutant glycan, which has the completely formed core structure and lack just the distal xylose moiety, thus demonstrating the proposed activity. The very low availability of the mutant glycan prevented the possibility to go on in the enzymatic characterization. However, chemical synthesis of glycan acceptors

is currently underway in the laboratory of Todd Lowarty at the University of Alberta and it will provide important information about the minimum glycan composition required for recognition by A075L active site.

In conclusion, the results obtained in this work contribute to shed light on the complex glycosylation machinery of giant viruses, in particular of Chloroviruses. Evidences obtained so fare clearly indicate that the viral glycosylation machinery is unusual and different from the ones found in cellular organisms, but how it was established and evolved is not clear. Many important pieces are still missing to draw the complete picture: for instance it is still debated if glycan formation occurs by a sequential step-by–step elongation of the glycan already bound to the protein, as it happens for O-linked glycosylation in Eukaryotes, or if a complete glycan is formed on a lipid linked precursor and it is then transferred "en-bloc" to the protein, with a mechanism similar to the N-linked formation in both Eukaryotes and Bacteria.

For Chlorovirus no information is known about the enzyme responsible for the attachment of the β-glucose to the Asn lateral chain. This type of linkage is very uncommon and very few enzymes have been identified to be able to catalyse this reaction in Bacteria and Archaea[50]. However, no homologs for these sequences were identified in PBCV-1 and other Chlorovirus genomes. So, the search for this possible protein-N-glycosyltransferase is still ongoing. Thus, several questions are still open about the formation of viral glycoproteins, including the mechanisms for glycan production and their subcellular localization, and about the origin of the viral enzymes and their relationships with those from cellular organisms. The identification of the viral enzymes involved in these pathways represents the starting point to clarify all these issues[7].

The presence of a complex and partially well conserved glycosylation machinery in Chlorovirus suggest that surface glycans are highly important at some point of the virus replicative cycle. Indeed, the spontaneous mutants that show aberrant glycoforms identified so far have a strongly reduced virulence. The heavily glycosylated surface can in fact provide protection against the external environment, as it is seen for bacterial capsular material and for some types of

spores[7]. Moreover, the glycan presence can also help capsid protein folding and promote the stability of the viral particle.

The study of the giant virus glycosylation pathways is important since it can increase our understanding about their life cycle and evolution. Moreover, the viral glycogenes, and in particular their glycosyltransferases, could also have interesting biotechnological application. GT are used to date to synthetized glycosides of pharmaceutical and industrial interest: glycosylated molecules (such as antibodies or polysaccharides) are used as drugs for many pathologies, from cancer to vaccines[4,23]. Usually, the synthesis of new drugs is limited from the fact that GTs are really specific, and sometimes a variation in the glycan decoration is needed to modify the pharmacokinetic/pharmacodynamics of glycosides. At this point, protein engineering is applied in order to modify the specify of GTs for glycan substrates and acceptors.

The identification of new family of GTs, such as A064R D2 that belongs to a new class mechanism, could be of special interest. Moreover, D1 domain of A064R does not have a strict requirement of a complex glycan as acceptor and it could for instance useful for the in vitro production of rhamnolipids. Indeed, these compounds are presently of great interests for several applications, but, since they are mainly purified by Bacteria, such as Pseudomonas, their biomedical use is often prevented by the possible carryover of contaminants.

Other polysaccharides are directly purified from target organisms and can be difficult to obtain for the copurification of impurities or in term of yield and, as additional problem, the direct chemical synthesis could be tricky[22]. As the viral GTs displays high solubility and high production yield in *E. coli*, the possibility to exploit these enzymes for in vitro glycan synthesis should be considered and investigated.

4. Future perspectives

As it is largely described in the discussion section, the discover and the study of new classes of GTs have an important biotechnological interest. At this purpose, the understanding of the new enzymatic mechanism identified for A064R D2 is the next step that need to be investigated. A064R D2 does not match with any already identified domain in the databases, suggesting that it could represent a new GT fold: starting from this observation, the NCLDVs glycogenome might be analysed in order to identify similar domains or, on the contrary, putative viral GTs that does not have any matches could be investigated to explore new GTs mechanisms.

According to that, the full comprehension of PBCV-1 glyco system should be pursued. As it is described in the previous pages, there are still some PBCV-1 putative GTs that are not characterized yet, such as the multidomain enzyme A111/114R. The characterisation of these enzymes and the resolution of the GT domain structure, will be then helpful to classify new domains that are not identified yet in the viral glycogenome, not only in Chloroviruses, but all over the Megavirales order.

Bibliography

(1) Etten, J. L. V. The Recent Discovery of Really, Really Big Viruses Is Changing Views about the Nature of Viruses and the History of Life. 19.

(2) Van Etten, J. L.; Agarkova, I.; Dunigan, D. D.; Tonetti, M.; De Castro, C.; Duncan, G. A. Chloroviruses Have a Sweet Tooth. *Viruses* **2017**, *9* (4). https://doi.org/10.3390/v9040088.

(3) De Castro, C.; Speciale, I.; Duncan, G.; Dunigan, D. D.; Agarkova, I.; Lanzetta, R.; Sturiale, L.; Palmigiano, A.; Garozzo, D.; Molinaro, A.; et al. N-Linked Glycans of Chloroviruses Sharing a Core Architecture without Precedent. *Angew. Chem. Int. Ed.* **2016**, *55* (2), 654–658. https://doi.org/10.1002/anie.201509150.

(4) Krauth, C.; Fedoryshyn, M.; Schleberger, C.; Luzhetskyy, A.; Bechthold, A. Engineering a Function into a Glycosyltransferase. *Chemistry & Biology* **2009**, *16* (1), 28–35. https://doi.org/10.1016/j.chembiol.2008.12.003.

(5) Freeze, H. H.; Schachter, H. Genetic Disorders of Glycosylation. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Etzler, M. E., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 2009.

(6) Varki, A.; Esko, J. D.; Colley, K. J. Cellular Organization of Glycosylation. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Etzler, M. E., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 2009.

(7) Piacente, F. Glycobiology: Characterization of NCLDVs' Glycosylation Mechanisms.

(8) Moremen, K. W.; Tiemeyer, M.; Nairn, A. V. Vertebrate Protein Glycosylation: Diversity, Synthesis and Function. *Nat Rev Mol Cell Biol* **2012**, *13* (7), 448–462. https://doi.org/10.1038/nrm3383.

(9) Stanley, P.; Schachter, H.; Taniguchi, N. *N-Glycans*, 2nd ed.; Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Etzler, M. E., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 2009.

(10) Rini, J.; Esko, J.; Varki, A. Glycosyltransferases and Glycan-Processing Enzymes. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Etzler, M. E., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 2009.

(11) Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E. *Glycans in Physiology and Disease*; Cold Spring Harbor Laboratory Press, 2009.

(12) Chung, C.-Y.; Majewska, N. I.; Wang, Q.; Paul, J. T.; Betenbaugh, M. J. SnapShot: N-Glycosylation Processing Pathways across Kingdoms. *Cell* **2017**, *171* (1), 258-258.e1. https://doi.org/10.1016/j.cell.2017.09.014.

(13) Nothaft, H.; Szymanski, C. M. Protein Glycosylation in Bacteria: Sweeter than Ever. *Nat. Rev. Microbiol.* **2010**, *8* (11), 765–778. https://doi.org/10.1038/nrmicro2383.

(14) Tu, L.; Banfield, D. K. Localization of Golgi-Resident Glycosyltransferases. *Cell. Mol. Life Sci.* **2010**, *67* (1), 29–41. https://doi.org/10.1007/s00018-009-0126-z.

(15) Opat, A. S.; van Vliet, C.; Gleeson, P. A. Trafficking and Localisation of Resident Golgi Glycosylation Enzymes. *Biochimie* **2001**, *83* (8), 763–773. https://doi.org/10.1016/S0300-9084(01)01312-8.
(16) Hassinen, A.; Kellokumpu, S. Organizational Interplay of Golgi N-Glycosyltransferases Involves Organelle Microenvironment-Dependent Transitions between Enzyme Homo- and Heteromers. *J Biol Chem* **2014**, *289* (39), 26937–26948. https://doi.org/10.1074/jbc.M114.595058.
(17) Campbell, J. A.; Davies, G. J.; Bulone, V.; Henrissat, B. A Classification of Nucleotide-Diphospho-Sugar Glycosyltransferases Based on Amino Acid Sequence Similarities. *Biochem. J.* **1997**, *326 (Pt 3)*, 929–939. https://doi.org/10.1042/bj3260929u.
(18) CAZy - Home http://www.cazy.org/ (accessed Oct 20, 2019).
(19) Lairson, L. L.; Henrissat, B.; Davies, G. J.; Withers, S. G. Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* **2008**, *77*, 521–555. https://doi.org/10.1146/annurev.biochem.76.061005.092322.
(20) Luzhetskyy, A.; Méndez, C.; Salas, J. A.; Bechthold, A. Glycosyltransferases, Important Tools for Drug Design. *Curr Top Med Chem* **2008**, *8* (8), 680–709. https://doi.org/10.2174/156802608784221514.
(21) Williams, G. J.; Gantt, R. W.; Thorson, J. S. The Impact of Enzyme Engineering upon Natural Product Glycodiversification. *Curr Opin Chem Biol* **2008**, *12* (5), 556–564. https://doi.org/10.1016/j.cbpa.2008.07.013.
(22) Lee, C.-J.; Lee, L. H.; Lu, C.; Wu, A. Bacterial Polysaccharides as Vaccines — Immunity and Chemical Characterization. In *The Molecular Immunology of Complex Carbohydrates —2*; Wu, A. M., Ed.; Springer US: Boston, MA, 2001; Vol. 491, pp 453–471. https://doi.org/10.1007/978-1-4615-1267-7_30.
(23) Anish, C.; Schumann, B.; Pereira, C. L.; Seeberger, P. H. Chemical Biology Approaches to Designing Defined Carbohydrate Vaccines. *Chem. Biol.* **2014**, *21* (1), 38–50. https://doi.org/10.1016/j.chembiol.2014.01.002.
(24) Forterre, P. Defining Life: The Virus Viewpoint. *Orig Life Evol Biosph* **2010**, *40* (2), 151–160. https://doi.org/10.1007/s11084-010-9194-1.
(25) Yutin, N.; Koonin, E. V. Hidden Evolutionary Complexity of Nucleo-Cytoplasmic Large DNA Viruses of Eukaryotes. *Virol J* **2012**, *9* (1), 161. https://doi.org/10.1186/1743-422X-9-161.
(26) Iyer, L. M.; Balaji, S.; Koonin, E. V.; Aravind, L. Evolutionary Genomics of Nucleo-Cytoplasmic Large DNA Viruses. *Virus Research* **2006**, *117* (1), 156–184. https://doi.org/10.1016/j.virusres.2006.01.009.
(27) Koonin, E. V.; Yutin, N. Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. *Adv. Virus Res.* **2019**, *103*, 167–202. https://doi.org/10.1016/bs.aivir.2018.09.002.
(28) Wilson, W. H.; Van Etten, J. L.; Allen, M. J. The Phycodnaviridae: The Story of How Tiny Giants Rule the World. In *Lesser Known Large dsDNA Viruses*; Van Etten, J. L., Ed.; Compans, R. W., Cooper, M. D., Honjo, T., Koprowski, H., Melchers, F., Oldstone, M. B. A., Olsnes, S., Vogt, P. K., Series Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2009; Vol. 328, pp 1–42. https://doi.org/10.1007/978-3-540-68618-7_1.
(29) Abergel, C.; Legendre, M.; Claverie, J.-M. The Rapidly Expanding Universe of Giant Viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol Rev* **2015**, *39* (6), 779–796. https://doi.org/10.1093/femsre/fuv037.
(30) Okamoto, K.; Miyazaki, N.; Reddy, H. K. N.; Hantke, M. F.; Maia, F. R. N. C.; Larsson, D. S. D.; Abergel, C.; Claverie, J.-M.; Hajdu, J.; Murata, K.; et al. Cryo-EM Structure of a Marseilleviridae Virus Particle Reveals a Large Internal

Microassembly. *Virology* **2018**, *516*, 239–245. https://doi.org/10.1016/j.virol.2018.01.021.
(31) Jha, S.; Rollins, M. G.; Fuchs, G.; Procter, D. J.; Hall, E. A.; Cozzolino, K.; Sarnow, P.; Savas, J. N.; Walsh, D. Trans-Kingdom Mimicry Underlies Ribosome Customization by a Poxvirus Kinase. *Nature* **2017**, *546* (7660), 651–655. https://doi.org/10.1038/nature22814.
(32) Claverie, J. M.; Abergel, C.; Ogata, H. Mimivirus. *Curr. Top. Microbiol. Immunol.* **2009**, *328*, 89–121. https://doi.org/10.1007/978-3-540-68618-7_3.
(33) Fischer, M. G.; Suttle, C. A. A Virophage at the Origin of Large DNA Transposons. *Science* **2011**, *332* (6026), 231–234. https://doi.org/10.1126/science.1199412.
(34) Poxvirus DNA Replication https://cshperspectives.cshlp.org/content/5/9/a010199.full.pdf+html (accessed Jan 9, 2020).
(35) Tulman, E. R.; Delhon, G. A.; Ku, B. K.; Rock, D. L. African Swine Fever Virus. In *Lesser Known Large dsDNA Viruses*; Van Etten, J. L., Ed.; Current Topics in Microbiology and Immunology; Springer: Berlin, Heidelberg, 2009; pp 43–87. https://doi.org/10.1007/978-3-540-68618-7_2.
(36) Bigot, Y.; Renault, S.; Nicolas, J.; Moundras, C.; Demattei, M.-V.; Samain, S.; Bideshi, D. K.; Federici, B. A. Symbiotic Virus at the Evolutionary Intersection of Three Types of Large DNA Viruses; Iridoviruses, Ascoviruses, and Ichnoviruses. *PLoS One* **2009**, *4* (7). https://doi.org/10.1371/journal.pone.0006397.
(37) Fischer, M. G. Giant Viruses Come of Age. *Current Opinion in Microbiology* **2016**, *31*, 50–57. https://doi.org/10.1016/j.mib.2016.03.001.
(38) Claverie, J.-M.; Ogata, H.; Audic, S.; Abergel, C.; Suhre, K.; Fournier, P.-E. Mimivirus and the Emerging Concept of "Giant" Virus. *Virus Res.* **2006**, *117* (1), 133–144. https://doi.org/10.1016/j.virusres.2006.01.008.
(39) Piacente, F.; Gaglianone, M.; Laugieri, M.; Tonetti, M. The Autonomous Glycosylation of Large DNA Viruses. *IJMS* **2015**, *16* (12), 29315–29328. https://doi.org/10.3390/ijms161226169.
(40) Mutsafi, Y.; Zauberman, N.; Sabanay, I.; Minsky, A. Vaccinia-like Cytoplasmic Replication of the Giant Mimivirus. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107* (13), 5978–5982. https://doi.org/10.1073/pnas.0912737107.
(41) Milrot, E.; Shimoni, E.; Dadosh, T.; Rechav, K.; Unger, T.; Van Etten, J. L.; Minsky, A. Structural Studies Demonstrating a Bacteriophage-like Replication Cycle of the Eukaryote-Infecting Paramecium Bursaria Chlorella Virus-1. *PLoS Pathog* **2017**, *13* (8), e1006562. https://doi.org/10.1371/journal.ppat.1006562.
(42) Van Etten, J. L.; Burbank, D. E.; Meints, R. H. Replication of the Algal Virus PBCV-1 in UV-Irradiated Chlorella. *Intervirology* **1986**, *26* (1–2), 115–120. https://doi.org/10.1159/000149689.
(43) Koonin, E. V.; Senkevich, T. G.; Dolja, V. V. The Ancient Virus World and Evolution of Cells. *Biol. Direct* **2006**, *1*, 29. https://doi.org/10.1186/1745-6150-1-29.
(44) Colson, P.; De Lamballerie, X.; Yutin, N.; Asgari, S.; Bigot, Y.; Bideshi, D. K.; Cheng, X.-W.; Federici, B. A.; Van Etten, J. L.; Koonin, E. V.; et al. "Megavirales", a Proposed New Order for Eukaryotic Nucleocytoplasmic Large DNA Viruses. *Arch Virol* **2013**, *158* (12), 2517–2521. https://doi.org/10.1007/s00705-013-1768-6.
(45) Milrot, E.; Mutsafi, Y.; Fridmann-Sirkis, Y.; Shimoni, E.; Rechav, K.; Gurnon, J. R.; Van Etten, J. L.; Minsky, A. Virus-Host Interactions: Insights from the Replication Cycle of the Large *Paramecium Bursaria Chlorella Virus*: Replication Factories of PBCV-1. *Cell Microbiol* **2016**, *18* (1), 3–16. https://doi.org/10.1111/cmi.12486.
(46) de Castro, I. F.; Volonté, L.; Risco, C. Virus Factories: Biogenesis and Structural Design. *Cell. Microbiol.* **2013**, *15* (1), 24–34. https://doi.org/10.1111/cmi.12029.

(47) Van Etten, J. L.; Dunigan, D. D. Chloroviruses: Not Your Everyday Plant Virus. *Trends in Plant Science* **2012**, *17* (1), 1–8. https://doi.org/10.1016/j.tplants.2011.10.005.
(48) Brussaard, C. P. D. Viral Control of Phytoplankton Populations-a Review1. *J Eukaryotic Microbiology* **2004**, *51* (2), 125–138. https://doi.org/10.1111/j.1550-7408.2004.tb00537.x.
(49) De Castro, C.; Klose, T.; Speciale, I.; Lanzetta, R.; Molinaro, A.; Van Etten, J. L.; Rossmann, M. G. Structure of the Chlorovirus PBCV-1 Major Capsid Glycoprotein Determined by Combining Crystallographic and Carbohydrate Molecular Modeling Approaches. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115* (1), E44–E52. https://doi.org/10.1073/pnas.1613432115.
(50) De Castro, C.; Molinaro, A.; Piacente, F.; Gurnon, J. R.; Sturiale, L.; Palmigiano, A.; Lanzetta, R.; Parrilli, M.; Garozzo, D.; Tonetti, M. G.; et al. Structure of N-Linked Oligosaccharides Attached to Chlorovirus PBCV-1 Major Capsid Protein Reveals Unusual Class of Complex N-Glycans. *Proc Natl Acad Sci USA* **2013**, *110* (34), 13956–13960. https://doi.org/10.1073/pnas.1313005110.
(51) Speciale, I.; Duncan, G. A.; Unione, L.; Agarkova, I. V.; Garozzo, D.; Jimenez-Barbero, J.; Lin, S.; Lowary, T. L.; Molinaro, A.; Noel, E.; et al. The *N* -Glycan Structures of the Antigenic Variants of Chlorovirus PBCV-1 Major Capsid Protein Help to Identify the Virus-Encoded Glycosyltransferases. *J. Biol. Chem.* **2019**, *294* (14), 5688–5699. https://doi.org/10.1074/jbc.RA118.007182.
(52) Tonetti, M.; Zanardi, D.; Gurnon, J. R.; Fruscione, F.; Armirotti, A.; Damonte, G.; Sturla, L.; De Flora, A.; Van Etten, J. L. *Paramecium Bursaria Chlorella* Virus 1 Encodes Two Enzymes Involved in the Biosynthesis of GDP-L-Fucose and GDP-D-Rhamnose. *J. Biol. Chem.* **2003**, *278* (24), 21559–21565. https://doi.org/10.1074/jbc.M301543200.
(53) Kang, M.; Dunigan, D. D.; VAN Etten, J. L. Chlorovirus: A Genus of Phycodnaviridae That Infects Certain Chlorella-like Green Algae. *Mol. Plant Pathol.* **2005**, *6* (3), 213–224. https://doi.org/10.1111/j.1364-3703.2005.00281.x.
(54) Nissimov, J. I.; Pagarete, A.; Ma, F.; Cody, S.; Dunigan, D. D.; Kimmance, S. A.; Allen, M. J. Coccolithoviruses: A Review of Cross-Kingdom Genomic Thievery and Metabolic Thuggery. *Viruses* **2017**, *9* (3). https://doi.org/10.3390/v9030052.
(55) Wilson, W. H. Coccolithovirus-Emiliania Huxleyi Dynamics: An Introduction to the Coccolithovirocell. *Perspectives in Phycology* **2015**, 91–103. https://doi.org/10.1127/pip/2015/0032.
(56) Piacente, F.; De Castro, C.; Jeudy, S.; Gaglianone, M.; Laugieri, M. E.; Notaro, A.; Salis, A.; Damonte, G.; Abergel, C.; Tonetti, M. G. The Rare Sugar N-Acetylated Viosamine Is a Major Component of Mimivirus Fibers. *J. Biol. Chem.* **2017**, *292* (18), 7385–7394. https://doi.org/10.1074/jbc.M117.783217.
(57) Zhang, Y.; Xiang, Y.; Van Etten, J. L.; Rossmann, M. G. Structure and Function of a Chlorella Virus-Encoded Glycosyltransferase. *Structure* **2007**, *15* (9), 1031–1039. https://doi.org/10.1016/j.str.2007.07.006.
(58) Wang, I. N.; Li, Y.; Que, Q.; Bhattacharya, M.; Lane, L. C.; Chaney, W. G.; Van Etten, J. L. Evidence for Virus-Encoded Glycosylation Specificity. *Proc Natl Acad Sci U S A* **1993**, *90* (9), 3840–3844.
(59) McPherson, A.; Gavira, J. A. Introduction to Protein Crystallization. *Acta Crystallogr F Struct Biol Commun* **2013**, *70* (Pt 1), 2–20. https://doi.org/10.1107/S2053230X13033141.
(60) PCT test https://hamptonresearch.com/product-PCT-Pre-Crystallization-Test-10.html (accessed Aug 20, 2019).

(61) Kabsch, W. Integration, Scaling, Space-Group Assignment and Post-Refinement. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (Pt 2), 133–144. https://doi.org/10.1107/S0907444909047374.
(62) Ley, R. E. Gut Microbiota in 2015: Prevotella in the Gut: Choose Carefully. *Nat Rev Gastroenterol Hepatol* **2016**, *13* (2), 69–70. https://doi.org/10.1038/nrgastro.2016.4.
(63) Busse-Wicher, M.; Wicher, K. B.; Kusche-Gullberg, M. The Exostosin Family: Proteins with Many Functions. *Matrix Biol.* **2014**, *35*, 25–33. https://doi.org/10.1016/j.matbio.2013.10.001.
(64) Barton, W. A.; Tzvetkova-Robev, D.; Erdjument-Bromage, H.; Tempst, P.; Nikolov, D. B. Highly Efficient Selenomethionine Labeling of Recombinant Proteins Produced in Mammalian Cells. *Protein Sci* **2006**, *15* (8), 2008–2013. https://doi.org/10.1110/ps.062244206.
(65) Iyer, L. M.; Aravind, L.; Koonin, E. V. Common Origin of Four Diverse Families of Large Eukaryotic DNA Viruses. *J Virol* **2001**, *75* (23), 11720–11734. https://doi.org/10.1128/JVI.75.23.11720-11734.2001.

Thanks to all of you!

This work was also possible thanks to the collaboration with Professor Jesus Jimènez Barbero and Doctor Adriana Rojas from CiC Biogune in Bilbao (Spain), and also thanks to Professor Cristina De Castro and Doctor Immacolata Speciale at the University Federico II in Naples, Italy.
At CiC Biogune I spent 3 months of my phD and here all the crystallization experiments, ITC experiment and the A075L NMR experiments took place. Here I I had the opportunity to learn new technics and amply my scientific skills.
The experiments in collaboration with Professor Cristina De Castro and Doctor Immacolata Speciale, were of fundamental importance for the A064R NMR characterisation.
I would like to thank all the scientists that have participate in this project, giving me the possibility to create a great work. I am really satisfied with this thesis and with these three years, that allowed me to grow as a scientist and, first of all, as a Biochemist.
I will start to thank my tutor, Professor Michela Tonetti, that I worked with for 6 years. She supported me for all this time, and she taught me all my scientific knowledge.
Professor Antonio De Flora, who was a nice pleasure to have a lot of “science discussions”, will be always an inspiration for me.
Professor Cristina De Castro and Doctor Immacolata Speciale, which actively participate in the experiments helping and contributing with great results.
Doctor Adriana Rojas, the Chrystal Platform manager at CiC Biogune, that flanked me in all my A075L experiments teaching me a lot. I would also to thank all the Aitor Hierro team, with I had the opportunity to have a great friendship during my stay at CiC Biogune.
Professor Jesus Jimènez Barbero that gave me the possibility to work in his laboratory and at CiC Biogune.
And I will also thank my “minor” favourite scientists Alice Franchi Biagi and Matteo Gaglianone, that are part of my life and my days as my best friend and my future husband.
I would also to thank Paola, not a scientist but my friend from 25 years, always interested in my successes and my progresses.
And finally, I would like to thank Thriss, my really favourite scientist by hearth (and not by academic studies), my absolutely best friend that supported me for each tear and each smile of these three years.
Of course, I will also thank my family, that supported my studies and my dreams, not only for these three years, but for all my life.